Prioritizing genes in RNA_Seq expression analysis using the consensus from multiple approaches

A Major Project dissertation submitted

In partial fulfillment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Payal Jain

(DTU /13/M.TECH448)
Delhi Technological University, Delhi, India

Under the supervision of

Dr. Navneeta Bharadvaja

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, INDIA

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **“Prioritizing genes in RNA_Seq expression analysis using the consensus from multiple approaches”**, submitted by **PAYAL JAIN (2K13/BIO/19)** in partial fulfillment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by her under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Project Guide
(Dr. Navneeta Bharadvaja)
Department of Bio-Technology
Delhi Technological University

Head of Department
(Professor D. Kumar)
Department of Bio-Technology
Delhi Technological University

DECLARATION

I, Payal Jain, hereby declare that the project work entitles “**Prioritizing genes in RNA_Seq expression analysis using the consensus from multiple approaches**” has been carried out by me under the guidance of Dr. Navneeta Bharadvaja, in Delhi Technological University, New Delhi.

This dissertation is part of fulfillment of requirement for the degree of M.Tech in Bioinformatics. This is the original work and has not been submitted for any other degree in any other university.

Payal Jain

(2K13/BIO/19)

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely fortunate to have got this all along the completion of our project work.

This thesis was made possible due to the masterly guidance of our mentor Dr. Navneeta Bharadvaja, who has the attitude and the substance of a genius: she continually and convincingly conveyed a spirit of adventure in regard to research and scholarship, and an excitement in regard to teaching. Without her guidance and persistent help this dissertation would not have been possible. We are grateful to her being very kind and patient while suggesting us the outlines of this project and correcting our doubts, for her constant supervision and support in completing the project.

We owe a debt of deepest gratitude to our Head of the Department, Biotechnology, Professor D. Kumar for providing us the necessary opportunities for the completion of our project.

We are indeed immensely grateful to (Dr.Shweta Choudhry, Dr.Guru Prasad, Mr.Deepak Kumar Shukla and Mr.Nilesh Kumar) whose contribution in stimulating suggestion helped us to coordinate the project and who truly provided us with the possibility to complete this project. A special thanks to Dr.Eugene Demidenko, Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, USA for allowing us to use his code and clarifying doubts in implementing his version of Gene Rank algorithm.

Heartfelt thanks to the other staff members of our department for their invaluable help and guidance.

PAYAL JAIN
2K13/BIO/19

CONTENTS

LIST OF FIGURES AND TABLE

LIST OF ABBREVEATIONS

AF	Activating Factor
AKT	Protein kinase B
BRCA	Breast Cancer
ER	Estrogen receptor
ERK	Extracellular signal regulated kinase
GO	Gene Ontology
GPER	G- coupled protein receptor
GR	Gene Rank
HER-2	Human epidermal growth factor receptor
MEK	Mitogen activated protein kinase
mTOR	Mammalian target of rapamycin
PCC	Pearson Correlation Coefficient
PI3K	Phosphoinositol-3- kinase
PR	Page Rank
Raf	Rapidly accelerated Fibrosarcoma
Ras	Rat Sarcoma
RNA_Seq	Ribonucleic Acid Sequencing
RPKM	Reads Per Kilo-base of exon model per Million mapped reads
RSEM	RNA_Seq by Expectation Maximization
TCGA	The Cancer Genome Atlas
TNBC	Triple Negative Breast Cancer

Prioritizing genes in RNA_Seq expression analysis using the consensus from multiple approaches

Payal Jain

Delhi Technological University, Delhi, India

1. ABSTRACT

Gene expression data gives us the knowledge of total mRNA molecules in a given sample. It can be measured using various techniques: serial analysis of gene expression, northern blots, microarrays, Reverse-transcriptase polymerase chain reaction, expressed sequence tag, Ribonucleic acid Sequencing (RNA_Seq) technology, Massively parallel signature sequencing, etc. high throughput technologies provide a great revolution in this vision. RNA_Seq process gains importance due to its effective and cheap sequencing. This technology is greatly used by the researchers in genomics. The Cancer Genome Atlas (TCGA) has used this approach for tumor analysis. In case of RNA_Seq data, gene expression can be quantified using Reads per Kilo-base of exon model per Million mapped reads (RPKM). Now-a-days, Breast cancer is more prevalent in women's causing to death. It is heterogeneous diseases, invasive or non-invasive in manner and categorized in hormone receptor-positive or triple-negative. The receptors can be human epidermal growth factors, hormone receptors (oestrogen and progesterone). In its signalling pathways, various gene are involved, to prioritize the gene for analysis by researchers various techniques are used. A very easy way to discovering interesting gene is comparison of expression profile of differentially expressed genes. Various approaches are available to extract the information from existing data using statistical methods. It can be Correlation coefficient method, Gene Rank, and Clustering. Correlation coefficient shows the linear relationship between two variables and their way of direction. It curtails the dimensionality of system. Gene Rank (GR) gives the ranking of gene in a given sample using Google Page Rank's (PR) algorithm. Clustering tells the genes which are more correlated to each other comes under same cluster and different genes in different clusters, this separation can be done on the pattern similarity basis. From this, we found that all three techniques in combination can be used to make a decision for predicting the gene priority and can be used for further analytical advancements.

2. INTRODUCTION

Genomics is the fundamental of cancer. The aim for studying genome sequencing of cancer cell may be any of them, to gather information about phenotype of cancerous cell, basic cancer biology and its treatment discovery. High throughput technology has tremendously increases the areas in research field, first genome sequencing after human genome project was done in 2008, after that in span of four years about 800 genomes has been sequenced for at least 25 different types of cancer (Mwenifumbo *et al.*, 2013). TCGA projects have used RNA_Seq approach to outline the primary tumor samples. The data is provided in four levels by different platforms. In data level 3, aggregate of processed data from single sample and is segmented, created by Reads per Kilo-base of exon model per Million mapped reads (RPKM) to gives the level of gene expression. It is one of the normalization methods to ensure the results of gene expression data (Mortazavi *et al*., 2008).

Cancer can be acquired (somatic) and inherited (genetic). The commencement and progression of breast cancer has been perceived as a secondary to the accumulation of genetic mutation which head to the aberrant cellular function. These mutations are either sporadic or inherited. It may inactivate the tumor suppressor genes and activates the onco-genes. It is heterogeneous diseases under various viewpoints (Bertucci *et al*., *2008*). There are various types of breast cancer, different in their diagnosis and prognosis: Luminal A and B, Her 2-positive, Basal-like (triple-negative) (Carol A. *et al*., 2014).

The rapid growth of molecular data in cancer enables complete and similar interpretation of heterogeneous genes linked to the traits in cancer (Vogelstein *et al.*, 2004). Statistical computation becomes inevitable in the field of advanced analytics of such data.

R-Statistical software is a freely available GNU project. It includes statistical algorithm, computation with annotation database and interactive visualizations. Its packages provide analysis of genomic data and equipped tools emboss analysis of data from sequencing methods. Various approaches for prioritizing the gene in RNA_Seq gene expression data analysis are:

Pearson correlation coefficient (PCC), measures the linear relationship between two continuous random variables. It is extensively used in similarity measure for gene expression data and based on pattern similarity check (Jiang *et al.*, 2004). It gives the strength and directionality of the relationship between genes.

Clustering is used to cluster or grouping of the similar gene separated from the dissimilar. There are various methods available for clustering, depending on different algorithm (Yeung *et al.*, 2001). Basically it is categorized into two methods: hierarchical and partitioning based.

Gene rank (GR) gives order of the important gene in an experiment based on phenotype and connectivity. It is based on Google's Page Rank algorithm concept. Microarray enriched gene rank monitors gene connectivity regardless of its phenotype nature (*E.* Demidenko, 2015).

Using existing data is still a new aspect in biology. The aim of our study is to show how information can be retrieved from publically available gene expression data and acts as a great utility factor for further analytical purpose. Here, to prioritize gene from RNA_Seq expression data of Breast cancer, we used the data from The Cancer Genomics Atlas portal. Different approaches have been used: Pearson Correlation coefficient, Clustering, and Gene Rank for making consensus to prioritize the genes in given data.

3. REVIEW OF LITERATURE

3.1 BREAST CANCER

Breast Cancer is the most prevalentcancer among the women in America, other than skin cancer, and second pre-eminent cause of death outstripped only by lung cancer. National Cancer Institute, estimates 232,340 female and 2,240 male breast cancers in the USA each year, as well as about 39,620 deaths caused by the disease. In 2015 United States, The American Cancer Society's reviews for breast cancer are: the recent cases of invasive breast cancer diagnosed will be 231,840 and non- invasive carcinoma in situ will be 60,290. The women will die from this cancer is about 40,290. According to National Cancer Institute's SEER database, In case of stage 0, the 5-year Relative Survival rate is 100%, stage I is 100%, stage II is 93%, stage III is 72%, stage IV is 22% is based on the prior version of AJCC staging.

Breast cancer is a heterogeneous nature disease (Bertucci *et al.*, 2008). It may be due to its diversified morphological features, clinical impact and response to therapeutic options (Viale *et al.*, 2012). High throughput technology gives more insight to inter-tumor and intra- tumor heterogeneity. Staging is to be done to know the intensity of breast cancer based on number of lymph nodes, size of tumor, invasive and non-invasive. Stage 0 includes ductal and lobular carcinoma in situ. Stage I shows tumor is about 2 cm or less and not spread to distant areas. Stage II larger than 5 cm but not spread to distant sites. Stage III shows tumor is of any size but not spread to distant nodal sites. Stage IV, tumor is of any size and spread to distant nodal sites. There is another TNM staging system which denotes size of the tumor, involvement of lymph node and metastasis.

There are various distinguish types of breast cancer: non-invasive ductal and invasive carcinoma, triple negative, inflammatory and metastatic breast cancer. The cancer that is associated with only glands and ducts of the breast is said to be non-invasive breast cancer that is different from benign. Another that spread to other surrounding tissues is called invasive breast cancer which is different from the secondary breast cancer. The cancer in which hormone epidermal growth factor receptor 2 (HER-2), estrogen receptors (ER), and progesterone receptors (PR) are not present is termed as triple negative breast cancer. Inflammatory breast cancer is fast growing in which cancer cells access the skin and lymph vessels of the breast, its symptom appears when lymph vessels get blocked. In metastasis, cancer cells invade to other parts of body through blood vessels.

Risk factors for the breast cancer are age and gender, family history, genes, menstrual cycle, alcohol use, child birth, diethylstilbestrol, obesity, hormone therapy and radiations.

The symptoms include: lumps in armpit, change in size and shape of breast, skin ulcers, weight loss, bone-pain. Various test used for diagnosis are breast exam, mammogram, breast ultrasound, biopsy, blood chemistry studies.

On the basis of cell or protein it can be categorized in hormone receptor-positive or triple-negative. With the help of biomarkers, such as hormone receptors (HRs) and human epidermal growth factor receptor-2 (HER2), breast cancer patients can be categorized into various subgroups with specific targeted treatment strategies. Approximately 75% of breast cancer shows oestrogen receptor. These are influenced to endocrine therapies in which it blocks and interfere with oestrogen receptor signalling. It transcriptional activity regulates PR expression.HER2 positive is responsible for about 15-20% of breast carcinoma.

In oestrogen signalling pathway (Wu *et al.,* 2015), oestrogen stimulates the cell proliferation through oestrogen receptor which involves ERα, ERβ, G-coupled protein receptor (GPER) and GPR-30.ERα plays important role in breast cancer malignancy. It encodes ligand which is dependent on nuclear receptor. It contain two domains that bind to de-oxy ribonucleic acid, Activating factor 1, 2 (AF1, 2) and one domain that binds to de-oxy ribonucleic acid. AF1 is not dependent on oestrogen receptor and work through growth factors by doing its phosphorylation. ERα not only confined to nuclear but also in cytoplasm, plasma membrane and mitochondria. Its domain AF-2 acts as a hotspot for the point mutation in breast cancer malignancy. It can function in both genomic and non-genomic action. In case of genomic, oestrogen first activates the ERα and it forms dimerized and then translocate it into the nucleus where it binds to the element of oestrogen receptor in genes. Further it activates or inhibits the gene, to which it binds. In oestrogen independent manner, ERα transcriptional activity is regulated by interaction with the co-repressors and co-activators. Specific protein-1 and activating protein-1 also binds to the ERα. The targeted genes for ERα are components of cell cycle, transcription and growth factors, ER. In non- genomic action, to regulate the cell proliferation ERα interacts with the signalling component outside the nucleus. In tissues, ERβ co-expresses with ERα. It surpasses the ERα function in cell proliferation. GPER binds to oestrogen and do some primary oestrogen signalling events present in plasma membrane.

In HER-2 signalling pathway, its output deviates in two axes of signalling. One is in from phosphatidylinositol-3-kinase (PI3K) to (AKT) to mammalian target of rapamycin pathway (mTOR) and another in this manner: from (RAF) to (RAS) then leads to (MEK) to (ERK). These signalling pathways play a vital role in cell proliferation, in its survival and metabolism, protein synthesis.

PI3K composes of regulatory and catalytic unit. Regulatory unit is p85 and three forms of catalytic unit of p110 are present: p110 α, β, δ. When-ever Her1-4 family of receptor kinase interacts with the PI3K it phosphorylates serione/ threonine kinase of AKT, then it targets to the ser/thr kinase of mTOR. Phosphatase and TENsin homologs can repress the phosphorylation of AKT. This gene is present in chromosome-10 called PTEM, a tumor suppressor gene. If it is lost, it creates malignancy in the breast cancer.

In RAS/ RAF/ MEK pathway, RAS activation forms hetero-dimer and activates RAF that further transmits signal to MEK-1 and MEK-2 and activates ERK-1 and ERK-2 which translocate into the nucleus. It initiates the process of transcription in dys-regulated manner in cell cycle progression and invasion.

Triple-negative breast cancer (TNBC) molecularly diverse (Lehmann *et al.*, 2015), serves as a collection of malignant breast tumors that have an increased risk of metastasis. Metastasis is the major reason of cancer-related deaths, including those in TNBC, and the presence of dormant residual disseminated tumor cells may be a key factor leading to metastasis (Chen *et al.,* 2015). It is defined by absence of oestrogen receptor, progesterone receptor and epidermal growth factor receptor (HER2/neu). It shows relapse pattern which is different from hormone positive cancer. Some overexpress epidermal growth factor receptor and some trans-membrane glycoprotein (NMB) (Anders *et al.,* 2008). It can be associated with an increased risk to nurture a BRCA1 mutation. Individuals with triple negative breast cancer also are at risk for a number of other germ line mutations, including mutations in the PALB2, CHEK2, BARD1, ATM, PTEN, BRCA2, and TP53 genes (Heikkinen *et al*., 2009; O'Brien *et al.*, 2014 and Churpek *et al*., 2015).

The therapeutic choice for invasive and non-invasive cancer are varied and complex. It can be treated via surgery, radiation therapy, chemotherapy, hormone therapy, targeted, bone-directed therapy. It can be categorized on the grounds of at what stage they are used and how it works. In local therapy, tumor is confined to breast only so surgery or radiation therapy can be given but in systemic, targeted, chemotherapy and hormone therapy is to be prescribed. In adjuvant therapy, even after surgery it appears back so, both systemic and radiation-therapy is given, one at a time. In non- adjuvant, chemotherapy and hormone therapy is administered before surgery so that after surgery, no tumor appears.

The number and nature of genetic variants that predispose women to breast cancer interplay between those variants and environmental factors. The most commonly known genes in breast cancer are BRCA1 and BRCA2 (BReast Cancer genes 1 and 2). These genes are present in everyone but some are having mutation which increases the risk of breast cancer. A number of genes are known to be involved in inherited susceptibility to breast. The wide variety of work has been done on genes which are involved in DNA repair and single nucleotide polymorphisms (SNPs) associated with an increased risk of breast cancer.

3.2 GENE EXPRESSION DATA

Gene expression data features the absolute or relative plenty of mRNA molecules in a given biological sample. To measure expression data in order to quantitate number of mRNA molecules of species is rarely possible, in reality. There are various technologies to analyse the gene expression data such as serial analysis of gene expression, northern blots, microarrays, Reverse-transcriptase polymerase chain reaction, expressed sequence tag, RNA_Seq technology. Of all these technology, RNA_Seq are most high throughput and widely used. RNA sequencing (RNA_Seq) is a high-throughput technology that was newly developed in 2008 for comprehensive transcriptome study (Wang *et al*., 2009). The Cancer Genome Atlas and Encyclopedia of the regulatory elements projects have used RNA_Seq approach to outline the primary tumor samples and cell lines respectively (TCGA: data portal, ENCODE: data matrix, 2013). After this, RNA_Seq technology gained significances.

In order to understand the principle behind different algorithm of RNA_Seq gene expression data, we need to understand the basic principles on which RNA_Seq technology is based.

RNA_Seq technology is based on high throughput sequencing principle, gives single base resolution. Its uses expand deep sequencing technology. In some cases, it relies on genomic sequence and having low background noise. A complete or fractionated portion of RNA is converted into fragments of cDNA library with ligated adapters to one or both ends. It is then sequences with or without amplified from one or both ends. The number of reads comes depends on DNA-sequencing technology used, typically 30-400 base-pairs. After sequencing, these reads are then aligned to either the reference genome or de-novo, without genome sequence to produce a map, consists of transcriptional structure and gene expression level. There are four commercial next-generation sequencing (NGS) platforms available for RNA_Seq: Illumina, SOLID, Ion Torrent, and Roche 454. It concurrently maps transcribed regions and gene expression. It is having ability to distinguish different isoforms and also allelic expression. It also provides information about SNP in transcribed regions(Cloonan *et al*., 2008 and Morin *et al*., 2008).To quantify gene expression level, it shows dynamic range more than 9000 fold (Nagalakshmi *et al*., *2008*). It depends on low amount of RNA and relatively low cost for mapping the transcriptomes of large genomes. It also gives information regarding the functional pathways of gene and its regulation process (Khatoon *et al*., 2014). It can be used to detect variants in the sequence and through transcription it imparts an investigation of the basic tumor DNA sequence (Xu *et al*., 2013). For genome wide identification of germinline variants (Miller and Hill, 2013) and somatic mutations (Chandrasekharappa *et al*., 2013and shah *et al.*, 2009), some researchers has used RNA_Seq alone.

In 2006, National Cancer Institute and National Human Genome Research Institute launched the program named The Cancer Genome Atlas (TCGA). RNA_Seq derived data is one of the source of gene expression, collected by TCGA. It compiles and analyse the tumor samples and presents the information regarding participant in the program, metadata histopathology slide images and its molecular information. In this, every platform produces various data types, associated with data levels. In Illumina’s platform, the brief overview of steps involved in this is firstly library is prepared for the interested sample and then sequencing is done to convert input RNA into small DNA fragments. In library preparation, it includes ploy-A-RNA isolation, RNA fragmentation, RT to cDNA using random primer, adapter ligation, and size selection from gel and PCR enrichment. Then this cDNA library is placed in flow cell where amplification is done and then converted into double stranded DNA clusters. This flow cell is then putted into the sequencing machine, and sequencing is to be done in parallel. The given number of cycle is headed in which four fluorescently labelled nucleotides are added and emitted signals are recorded. This intensity then converted into base calls. The length of reads and its number is determined by the number of cycles and number of clusters respectively. Data levels characterize the data in TCGA so that researchers can locate their data of interest. Four levels of data are present in this: Data level 1 shows raw data, which is not normalized and low level for single sample. Level 2 shows normalized, processed data which is interpreted for the molecular abnormalities presence. Level 3 is an aggregate of processed data from single sample and is segmented or we can say an interpreted data. Level 4 shows for the particular region of interest, gives quantified association based on molecular abnormalities, clinical variables, sample’ characteristics. The data created in Level 3 is by two methods: the original method follows Reads per Kilo-base of exon model per Million mapped reads (RPKM) and another method is combination of Map_Slice and RNA_Seq by Expectation Maximization (RSEM) to determine the gene expression level.

In RNA_Seq analysis, sequenced reads have to be normalized. It is to be done to ensure the results of gene expression. It has been proposed to eliminate the unwanted variation in the data. There are various methods available for the normalization: Total counts, Upper quartiles, full quartiles, median, RPKM, boxplot of log 2, housekeeping gene counts. RPKM is one of those normalizing method, it first divides the sequenced reads by total library size and calculated as follows:

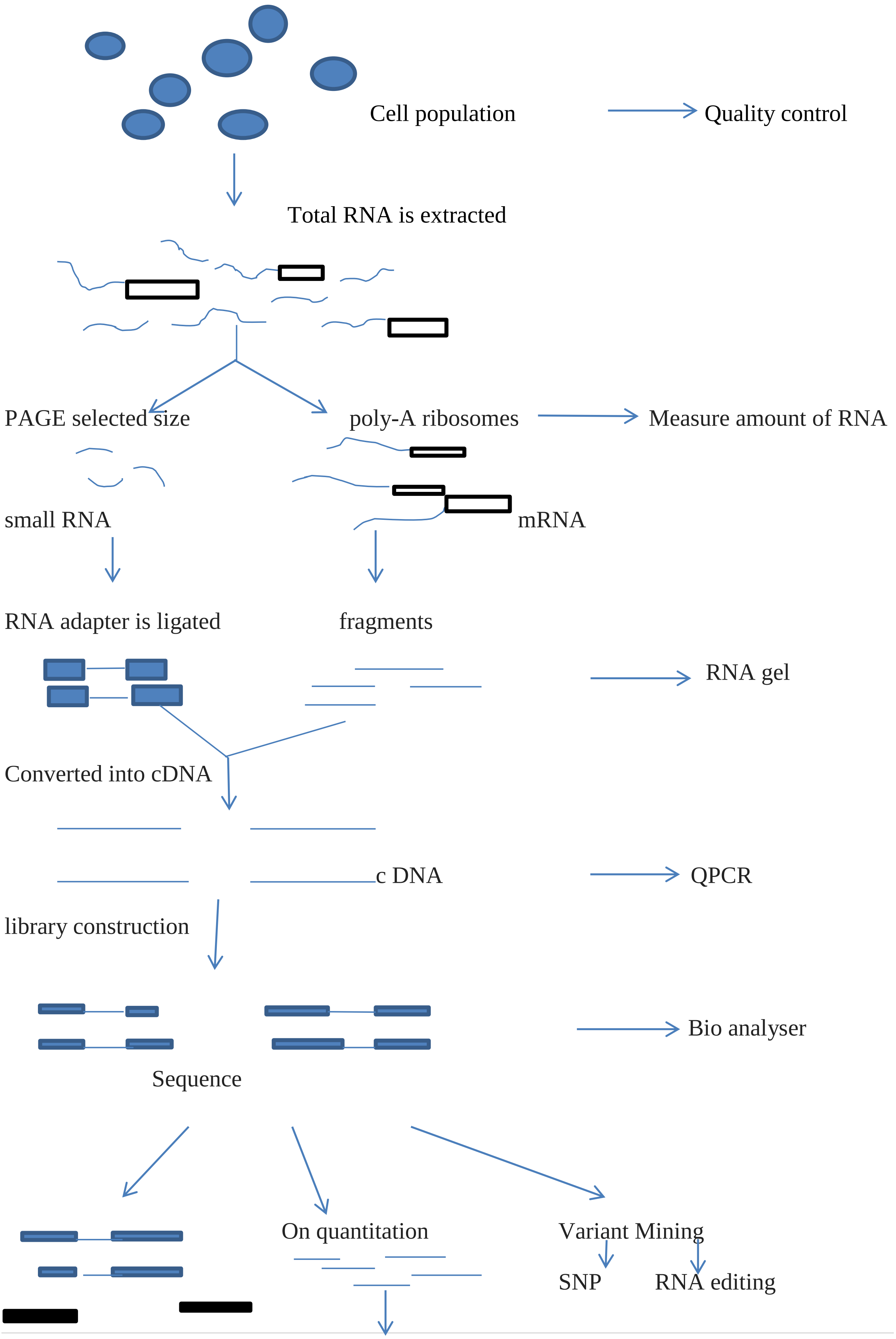
RPKM = 10^{9} *C / N*L

C denotes the number of reads that can be mapped onto the exon of genes.

N – Total number of reads in a sample.

L – Total sum of exon in base pairs.

RNA_Seq workflow



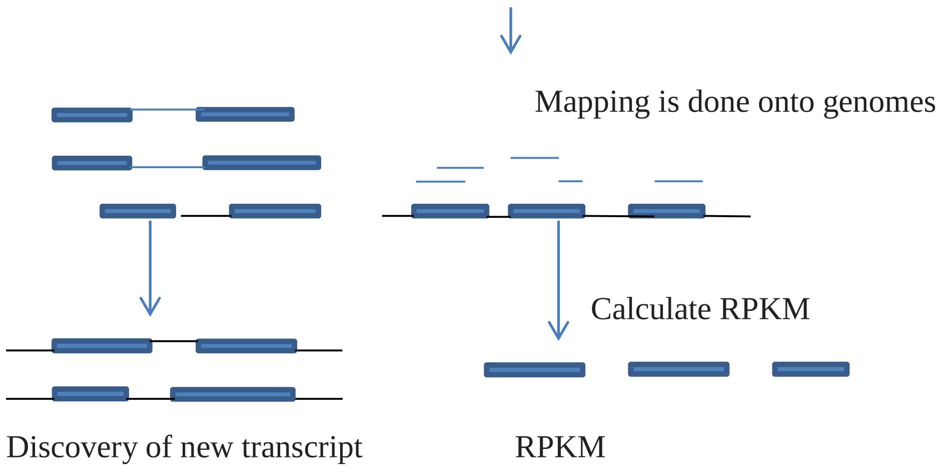


Fig 1: RNA_Seq workflow

Analysis of Gene expression matrix

The two ways to study the gene expression matrix are either comparing expression profiles of gene by comparing rows or sample by comparing column in gene expression matrix. This is done to look for either similarity or differences in the row or column. If rows are similar, then can be assumed the respective genes are co-regulated to each other or if columns are similar then genes are differentially expressed in samples. These objects are regarded as n-dimensional vector where n is the number of genes for sample comparison or a number of samples for gene comparison. Euclidean distance between object is the eventually choice for the comparison, but most apparent choice is to use correlation coefficient for gene expression data analysis.

We can study the expression matrix in either supervised or unsupervised way. The best example for unsupervised way is clustering where correlated genes or sample can be found. Clustering techniques are not new to analyse gene expression matrix but there are various algorithm have been proposed for it. A very easy way to discovering interesting gene is comparison of expression profile of differentially expressed genes (Brazma *et al*., 2000; 2003).

In analysing gene expression data, while extraction of information there are sometimes a problem occur that is, the data matrix showing some missing values or undefined in the gene expression data. There are wide variety of method are available to extract this information from the existing data using statistical methods. Row average method means which takes average of the interested gene of expression, weighted k-nearest neighbors and expectation maximization (Bo *et al*., 2004 and Troyanskaya *et al*., 2001).These data matrices are too large to study manually, so automated methods required to reduce the dimensionality.

3.2.1 PEARSON CORRELATION COEFFICIENT

In 1880's, correlation coefficient plan was introduced by the Francis Galton ,later on refined by the Karl Pearson (Galton *et al.*,1877,1885,1886; Pearson,1985 and Stigler *et al*.,1989). Pearson published the paper on treatment of correlation and regression in 1896 in the Philosophical Transactions of the Royal Society of London where, he credited *Bravais* for giving the initial correlation mathematical formulae in 1846 (Galton *et al*., 2001).

In the field of statistics, Correlation is a method to explore the relationship between the two quantitative, continuous variables. Pearson's correlation coefficient (PCC) is a measure of the strength of association between the two variables. It is extensively used in the sciences as a method of the degree of linear dependence between two variables. Its value lies between the range of +1 and-1, where if showing positive values then depicts that one variable is directly proportional to the other variable in positive manner and in case of negative values it shows variables are directly proportional but negatively correlated to each other and zero implies that there is no correlation between the variables. It tells how much two random variables vary together and divided by the product of their quantified amount of variation in a given set of values (standard deviation).

In Pearson product moment correlation method, suppose σ_a and σ_b are the standard deviations of two random variables a and b respectively. Then Pearson product moment correlation coefficient between the variables is denoted as:

$\rho_{a,b} = [cov(a,b)]= [E((a-E(a))(b-E(b))] / \sigma_{a*} \sigma_b$

E(.) shows the expected value of the variable, cov is covariance , σ_a is the standard deviation of variable a and σ_b is the standard deviation of variable b.

To implement this, we must be firm that the interval data attain from the paired observations and normally distributed variables. If data contains extreme values they may affect the results and sometimes it could be ambiguous when non-linear relationship variables are considered.

It can be calculated for population and sample represented by ρ and r respectively.

In case of population, Pearson correlation coefficient is described as:

$\rho_{a,b} = cov(a,b) / \sigma_{a*} \sigma_{b}$.

In terms of mean and expectation, cov (a,b)= E [(a- μ_a)(b- μ_b)]

It can also be: $\rho_{a,b} = E[(a- \mu_a)(b- \mu_b)] / \sigma_{a*} \sigma_b$, where μ_a is the mean of 'a' variable and μ_b is the mean of 'b' variable.

The formulae for $\mu_a = E$ (a), $\mu_b = E$ (b)

$\sigma^2_a = E [(a-E (a))^2] = E (a^2) - E (a)^2$

$\sigma^2_b = E [(b-E (b))^2] = E (b^2) - E (b)^2$

$E [(a- \mu_a) (b- \mu_b)] = E [(a-E (a)) (b-E (b))] = E (a*b)-E (a) * E (b)$

$\rho_{a,b} = E (a*b)-E (a) * E (b) / \surd E (a^2) - E (a)^2 * \surd E (b^2) - E (b)^2$

The research to analyse the sequenced data which are transformed into matrices of gene expression contains, genes in rows and sample in columns. PCC is extensively used in similarity measure for gene expression data and evinces effective in its analysis. It can be calculated corresponding to gene or sample depends on researcher's interest.

Proximity measurement measures the similarity or closeness strength between two data objects. In case of RNA_Seq data, gene behaves as an object where its connectivity is measured by the expression profile pattern of these genes. PCC method does the same by analysing the shape profile gene expression pattern and gives the correlated values (Jiang *et al.*, 2004). It accounts the rank of a gene expression variable. It also constructs the network of co-expressed genes. It is determined in the samples for all pairwise comparison of the values of gene expression. In (Butte *et al*., 1999), they studied Pearson correlated coefficient which was performed by a program written in MATLAB to find the network of related variables in medical datasets. They showed the network which was found consistent with basic human physiology.

Pearson correlation measure was transformed into a connection strength measure by using a power function (connection strength (i,j) = |correlation (i,j)|^β) (Zhang *et al.*, 2005). The dissimilarity of gene expression is based on 1- absolute value of PCC values (Bittner *et al.*, 2000).

3.2.2 CLUSTERING

RNA_Seq technology enables expression level measurement for thousands of gene in a parallel fashion, helping researchers to gather knowledge and insight about diverse biological phenomena. In order to unveil information contained in gene expression data, one of the first step usually adopted is cluster analysis, which finds its predominant application when genes that show similar expression patterns are clustered together. Clustering techniques is significant to explain the gene regulation, gene function and cellular processes.

It can be taken as a tool for reducing the system's dimensionality. In an experiment, to describe the state of tissue or cell, mean intensity of cluster of genes can be used in place of considering large number of gene intensity (U. Alon *et al.*, 1999). Cluster analysis is an unsupervised way of categorization of pattern into groups. It seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. In this, gene with similar pattern of expression is clustered together in same cluster depicts similar cellular function. In this way, it can be used to infer the information of gene which is previously undetermined (Eisen *et al.*, 1998). The hypothesis can be generated regarding gene transcriptional regulatory network mechanism by interpreting the regulation of gene through gene expression data cluster (D'haeseleer *et al.,* 1998). It is concise to cluster gene and sample both in the gene expression data. In case of gene based clustering, co-expressed genes occur in same cluster where gene is taken as an object and sample is feature. In sample based clustering, sample can be partitioned into homogeneous by taking sample as an object and gene as feature. There is no precise definition for it.

Basically, clustering methods have been divided into two types: hierarchical and partitioning based clustering. For making the evolutionary trees, Hierarchical clustering techniques,were shown to be valuable (Somogyi *et al.,* 1995; Spellman *et al.,* 1998 and Wen *et al.,* 1998). In Hierarchical based, it seeks to shape a hierarchy using agglomerative or divisive strategies. Its results are usually presented in a dendrogram manner. In partitioning based, it curtail a given clustering criterion by iteratively relocating data points between clusters until a locally partition is attained.

Its method is divided generally into two classes: supervised and unsupervised clustering. In supervised, gene vectors are categorized on the ground of its reference vector. In unsupervised classes, there is no or little prior knowledge of gene so it is done without any predefined reference vector.

There are many clustering methods which can be applied on the analysis of gene expression data. Some of them are Hierarchical clustering, Self-organizing maps, K-mean clustering, Methods based on within cluster maximization and between cluster similarity minimization, Ensemble method, Biclustering methods. Methods based on graph theory were used for the clustering of cDNAs based on their oligonucleotide (Hartuv *et al.,* 1999). K-means method is attributable to both simplicity and practicality. It supports the numeric columns. It is widely used for gene expression analysis in data mining. It is considerably faster than the iterative partition and Cluster Affinity Search Technique (CAST) algorithms.

Its aim is to make low variance within the clusters and large variant across the clusters by separating the objects into groups. Depending on mean expression vector in each cluster, distance between clusters can be evaluated (Parmigiani *et al.*, 2003).

K-means algorithm (k dhiraj *et al.,* 2009)

1. Take k initial cluster centre $\{S_1,S_2,...,S_k\}$ randomly from n-points$\{P_1,P_2,...,P_n\}$.
2. Assign points(P_a), where a =$\{1,2,\ldots,n\}$ to clusters $\{C_b\}$ where b=$\{1,2,\ldots,k\}$.
   If $\| P_a\text{-}S\| < \|P_a\text{-}S_c\|$ where c = $\{1,2,\ldots,k\}$ and b ~= c.
   Ties are resolved arbitrarily.
3. Compute new cluster centres$\{S_1^*,S_2^*,\ldots,S_k^*\}$as follows: $S_a=1/n \sum p_b \in C_aP_b$, where a=$\{1,2,..k\}$, n_a is the number of elements belonging to cluster C_a.
4. If S_a*=S_a, a =$\{1,2,\ldots,k\}$, then terminate the process here otherwise repeat step 2.

Flowchart of enhanced k-means clustering (Muhammad Rukunuddin *et al.*, 2013)

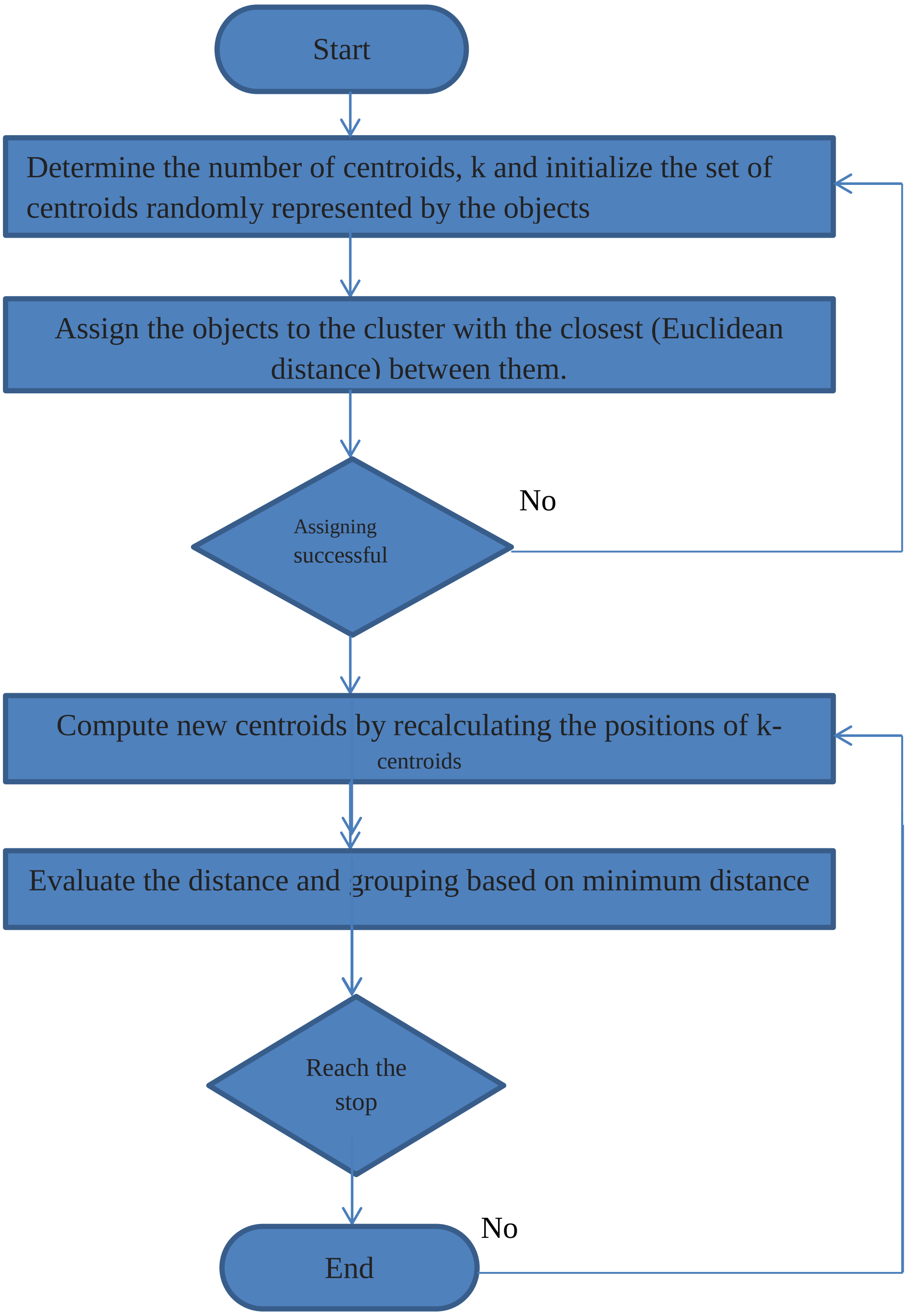

Start
Determine the number of centroids, k and initialize the set of centroids randomly represented by the objects
Assign the objects to the cluster with the closest (Euclidean distance) between them.
Assigning successful
No
Compute new centroids by recalculating the positions of k-centroids
Evaluate the distance and grouping based on minimum distance
Reach the stop
No
End

Fig 2: k-means algorithm flowchart in clustering methods.

3.2.3 GENE RANK

The Page Rank (PR) algorithm (Page *et al.*, 1999) is used by Google to rank the web pages in results of their search engine which appears to be similar to the situation of analysis of gene expression data. In case of PR, web pages are denoted as nodes, hyperlinks as edges. It shows directed graph network. The rank of page is given by sum of rank of those pages which are linked inbound to it and their rank is given by the rank of pages which links to them. It does not consider all the inbound links equally. Every page has some forward links (headed away from the page) and backlinks (towards the page) rank is assigned high if the sum of its backlinks is high. The Algorithm of PR has random walk interpretation which tells that, rank of webpage depends on time spent on page while surfing the web pages and also some other attributes such as page size, hyperlink and headline context, updation time of pages and number of changes done in it. Damping factor (d) is also known as decay factor, it shows the chance that user can get out of the current page by not clicking hyperlinks in it but request to random new page. The random walk distributed algorithm for nodes in a network takes O (log n /∊) for both directed and undirected network. N shows the network size and ∊ is the reset probability. In case of undirected graph PR is statistically near to the degree of graph Where F is the degree distribution vector

$F = 1 / 2|E|$ [deg (p_1), deg (p_2),..,deg (p_N)]

Then $1\text{-}d / 1\text{+}d||1 / N*1 \text{-}F||_1 \leq ||\ PR\text{-}F||_1 \leq ||1 / N*1 \text{-}F||_1$ shows page rank is equals to distribution vector if graph is regular.

Its mathematics is quite general and can be applied to any network or graph in any sphere. It can be computed in any of the two ways either algebraically or iterative method. The efficiency depends on frame work for the computation, definite implementation of methods and preciseness of results required; computation time can vary for these methods.

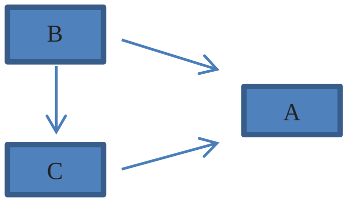


Fig 3: Page Rank basic concept of ranking.

Suppose A,B,C are the web pages, initially all pages are assigned same value, outbound links as L(B), L(C),then the PR for A is given by: PR (A) = PR (B) / 2 + PR (C) / 1 OR PR (B) / L (B) + PR (C) / L(C)

In general Page Rank, for any page x and y is expressed as:

$PR(x) = {}_{y \in B\ x} \sum PR(y) / L(y).$

PR of x dependent of PR value of y contained in set of B_x.

Page Rank of web page is updated when-ever any new page is added to it.

Gene rank (GR) is the immediate modification of the PR that preserves its mathematical properties. As in case of Google Page Rank, it counts the votes from highly ranked page, in the same way in case of gene expression data analysis, the expression measurements, such as protein interaction data, functional annotations, or previous experimental result is considered. In this, node is represented as a gene and edges as the previous knowledge. It shows un-directed network. Its algorithm provide enhanced ranking of the genes than the ranking of change of pure expressions and that can be further used in analytical advancements.

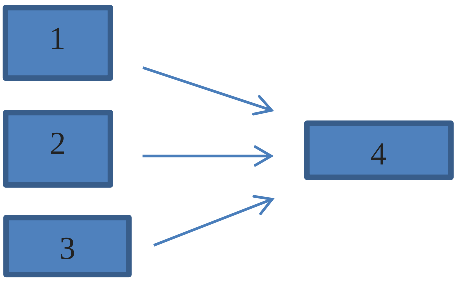


Fig 4: Gene rank concept.

Let us suppose, there are 4 genes, gene1, 2, 3 are showing highly differential gene expression value but gene 4 is depicts no differential expression, if it is a transcription factor that regulates the expression of those gene which are connected to it, so linked to highly differentiated gene make it ranking high in gene rank. This is the reason, why in RNA_Seq experiment a gene with lower in its position can get high rank.

Gene rank algorithm based on page rank algorithm (Morrison *et al.*, 2005).

Web Page Rank in PR algorithm is shown by (1-d), where d (damping factor) is taken as 0.85, so in case of gene rank it is considered as $(1\text{-d})_{ex}$ where ex is denoted as gene expression change values, and d (that controls the weighting of change in expression related to connectivity used in the measurements) is assumed to be 0.9. Suppose $p^{|n|}{}_{s}$ which shows the gene s, ranking after doing n number of iterations. Initially it is taken as $p^{[0]} = ex / \|ex\|_1$.

||.||1 shows vector 1-norm.

After that suppose $p^{|n|}{}_{s} = (1\text{-d})\ ex_{s} + d\ {}_{i=1}{}^{N}\sum W_{i\,s}\, p_{i}{}^{[n-1]} / \deg_{s}$, where $1 \leq s \leq N$.

$\deg_{s} := {}_{i=1}{}^{N}\sum\ W_{i\,s}$, W belongs to P^{N*N} shows gene network connectivity matrix, so $W_{is} = W_{si} = 1$ if both i and s gene are connected to each other, otherwise it will be 0. . If d=0, then p=ex which shows ranking is based only on gene expression level. If d=1, ranking is based on connectivity.

In case of PR algorithm transmit occur in the random walk process which is tendentious towards preferred location of user and in gene expression analysis, it is towards expression level.

The abstract principle of page rank algorithm has both random walk interpretation and vote of confidence and for the situation where ex $_i$ = 1, it reclaims the original Google Page rank

GR can be calculated using any of the two forms: Pearson Correlation coefficient or Gene Ontology (GO) annotation. PCC shows the extent and direction of the relationship between two variables. GO annotation appends the cellular component, molecular and biological process information to gene. If PCC denoted by r, $r > 0.5$ then gene are considered as connected which has no justification and correlation coefficient brushes off negative value. In case of GO annotation it is difficult to form a matrix with $2.5*10^9$ elements if 50,000 genes are being considered. Due to these drawbacks, ranking method is again revised by Eugene Demidenko.

Gene rank method (E. Demidenko, 2015) uses squared correlation coefficient (R^2) and this method is free from above shortcomings. Conventionally, GR is calculated using t-statistics or we can say that correlated with the phenotype (Winter *et al.*, 2012 *and* Zuber *et al.* 2009). But in this method, only gene connectivity problem is considered regardless of phenotype.

In this, data is expressed as n×m matrix where n denotes number of rows with gene and m is number of samples. Consider two gene s and i, PCC is calculated for n×n matrix gives both positive and negative values, which shows positive and negative relationships respectively. But here, only interest is in gene connectivity so squared correlation coefficient($R^2 = \{r^2_{si}\}$) is taken if it is more towards 0 shows no relationship, but if it is approaching 1 gives linear dependency. R^2 depicts the co-expression network. As r^2_{si} in the s th row shows its gene connectivity. This data is normalized so that gene in row can be compared to each other, which is given by:

$R^2{}_{*si} = r^2{}_{si} / {}^n{}_{k=1}\sum r^2{}_{sk}$

It is stated as normalized squared correlation matrix, belongs to the stochastic matrices where every element in the matrix is non-negative and its sum is 1.

Suppose p_s is the gene rank for the gene s, then another to calculate its weighted sum of squared correlation coefficient p_s: $\sum^n{}_{s=1} p_s r^2{}_{*si}$where, weighted is done with respect to connectivity and in page rank, there is iterative definition of p.

Then, iterative definition for microarray enriched gene rank:

$P_i = 1 / n * (1\text{-}ai) + a_i\ {}_{s=1}{}^n \sum p_s r^2{}_{*si}$; i=1,2,3….n

It represent earlier knowledge for the gene i connectivity and connectivity information from the expression data.

From this p=H’p where,

H= 1/ n (1-a) 1’ + AR*2

From the definition it can be analysed that the dissipation of sth gene rank to the remaining genes having n number of connections. And this connection can be represented by:

$M_{s\,i} = p_s\, p_i\, H_{i\,s} \times 100\%$.

All $M_{s\,i}$ are nonnegative and the sum of $M_{s\,i}$ over s = 1, 2, .., n is 100%.

Cluster analysis and gene rank

Cluster analysis techniques such as k-means and hierarchical is used to make grouping of genes showing some similarity pattern, before this data normalization is to be done. Pearson correlation coefficient can be expressed through calculating Euclidean distance between normalized samples of s and i genes. The formula is given by:

$Z_s - Z_i = \sqrt{2}\ (1-p_{si})$.

According to above formula, which gives the idea of there is a relationship exists between Gene rank and cluster analysis. From this, it can be interpreted that genes within same cluster showing genes with high gene rank due to their closeness and Gene Rank density, which is the mixture of number of components would be same the number of clusters

4. METHODOLOGY

In this, every step from file preparation to cluster was done in R.3.1.2 console window, code for which is given in appendix.

DATASET

In order to perform the analysis, Breast Cancer data was required. We collected all available tumor sample file data from TCGA, by applying filter settings for the disease-BRCA, (Breast Invasive Carcinoma), the platform UNC Illumina HiSeq_RNASeq level3 data (consists of exon quantification file contains calculated expression signal of gene, gene quantification file contains particular composite exon of gene’s expression signal, and in spljxn file, expression signal of particular composite splice junction of a gene).

The quality control and processing of the data were done by workgroup of Broad Institute’s TCGA. Trimmed annotated gene quantification files were selected for the analysis purpose.

FILE PREPARATION

In file preparation, retrieved only RPKM variable values of all the genes of data file and merged all the sample file data.

Merged data must ensure that it contains common genes from all the data file. So, we removed the redundant genes occurring in sample file. The genes containing zeros-value in the sample files were deleted.

GENE RANK EVALUATION

Gene rank was developed by Morrison *et al.*, based on the principle of PageRank in Google. There are two versions of Gene Rank: GO annotation and Pearson correlation coefficient. It shows the intricacy of genetic organisation. It shows new biological vision as connectivity within the group or cluster.

It was calculated by giving above merged data file as an input and then on using script gives the file containing gene with their rank value. In this, different damping factor values (a = 0.8, 0.85, and 0.9) has been used to plot the Gene rank density plot and rank value.

CALCULATING PEARSON CORRELATION COEFFICIENT

Correlation is a method to explore the relationship between the two quantitative, i.e. continuous variables. It measures the range to which two variables tends to change together. It shows both direction and strength of relationship.

Pearson’s correlation coefficient is a measure of the strength of association between the two variables. A relationship is linear in case of when change in one variable is proportional to the change in other variable.

It was done by giving above transpose merged file as an input and then run a command which results in genes correlation coefficient values with respect to each other.

CLUSTERING

Clustering is the fundamental task in data mining. Its model is used as both supervised and unsupervised learning method. Its goal is to find a new set of categories and their assessments are intrinsic. Many clustering methods are available with different induction principle. We used the K-means method under partitioning group.

It was performed by giving normalized matrix of merged file as an input and then after writing script based on k-means algorithm, it makes the required graph of cluster.

5. RESULTS

In this section, the results in tabular form are given in the DVD enclosed with it.

DATASET

The data file retrieved from TCGA (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) is 19.654GB. It contains RNA_Seq file folder consists of 881 sample file. The gene quantification file of 832MB consists of gene Entrez/LocusLink symbol followed by its ID, raw_counts, median length normalized and RPKM variables value of 20533genes.

FILE PREPARATION

In prepared file, we found only one redundant gene (SLC35E2) and 9 genes with zero-value in the all files, and 29 unknown genes which have been removed. The outcome file after merging all the sample files contains 20494 genes with their RPKM values forming 20494*881 matrix data file. The prepared file is shown in Table 1.

GENE RANK CALCULATION

Using the script for gene rank calculation, merged 20494*881 matrix data used as an input produces normalized matrix and then further processing gives the rank of all genes in output data file. As we have used different value for damping factor parameter, from all these we get almost similar rank of the gene with some variation in their values.

High ranking genes also involved in various other cancer types. The results are shown in Table 2, 3, 4 in which gene rank was calculated using different damping factor. In Table 2, we have considered damping factor, a= 0.80, Table 3 was calculated using a=0.85, Table 4 using a=0.90. Figure 5 shows gene rank density curve using a=0.8, Figure 6 shows with a=0.85 and Figure 7 shows gene rank density plot for a= 0.9.

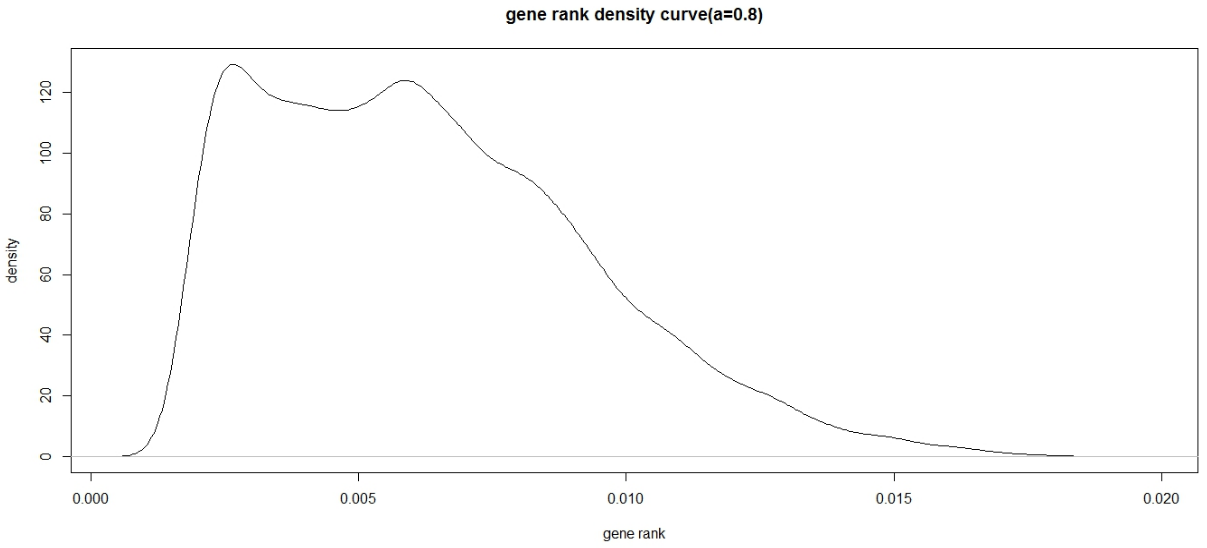


Fig 5: Gene rank density curve using a = 0.8 showing at the extreme rank values, density is low.

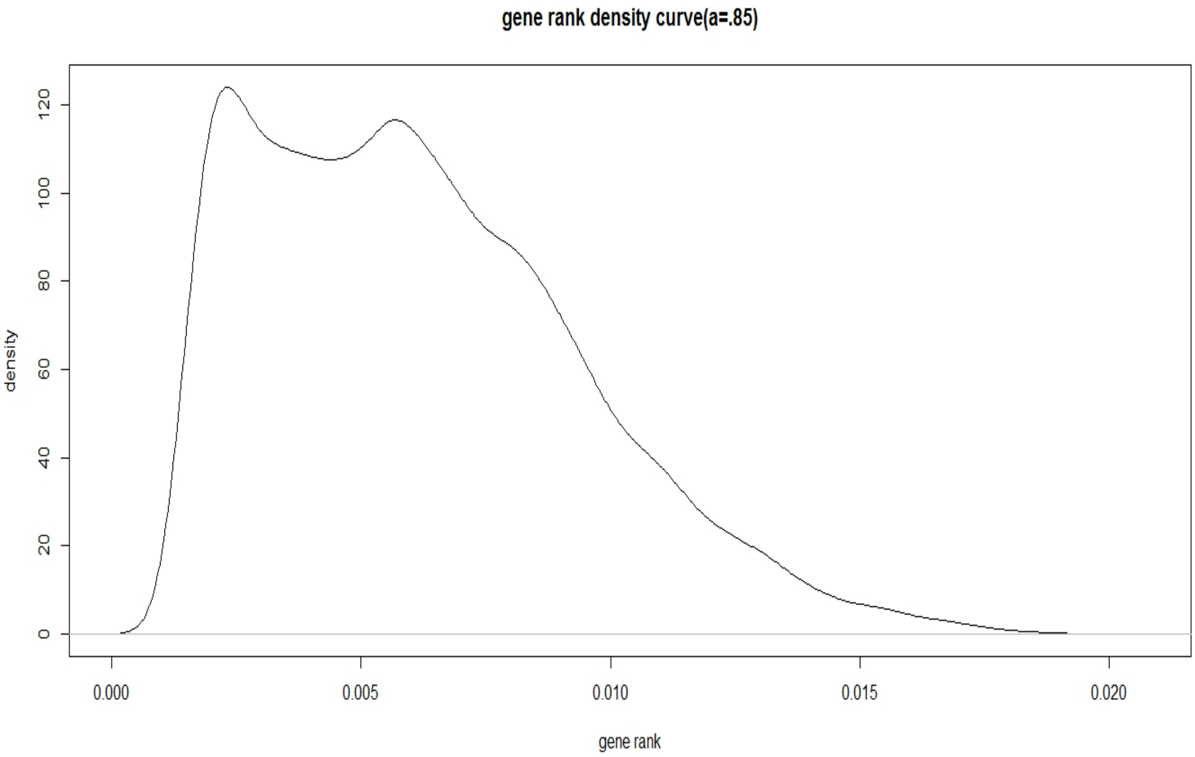


Fig 6: Gene rank density curve using a = 0.85

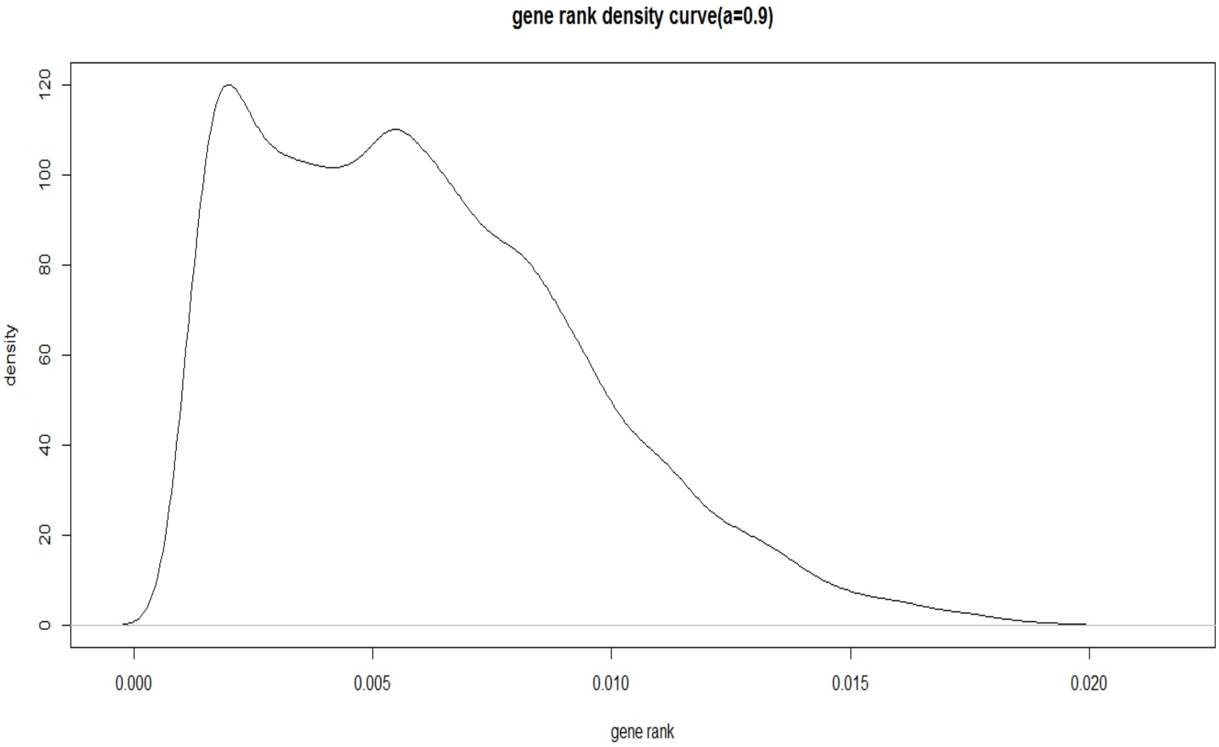


Fig 7: Gene rank density curve using a= 0.9

PEARSON CORRELATION COEFFICIENT

After transposition of above 20494*881 matrix data, gives 881*20494 matrix file. By giving this file as an input in R-console window and running a Pearson correlation coefficient script in, it gives 20494*20494 matrix data file which contains correlated coefficient values of gene with respect to each other.

According to gene rank calculation, different sets of higher ranking genes has been taken and then calculated their Pearson correlation coefficient for the analysis purpose. The genes which shows highly correlation to other gene is similar in their some process aspect. Results are displayed in the Table 5, 6, 7, 8, 9. Table5 shows the value of all gene expression data. In Table 6, we have taken top 50 ranking genes from calculated gene rank and then estimated their Pearson correlation coefficient values, Table 7 shows value for top 100 ranking genes, Table 8 shows value for top 300 ranking genes and Table 9 shows PCC value for top 500 ranking genes. Figure 8, 9, 10, 11 shows correlation plot of genes with respect to each other.

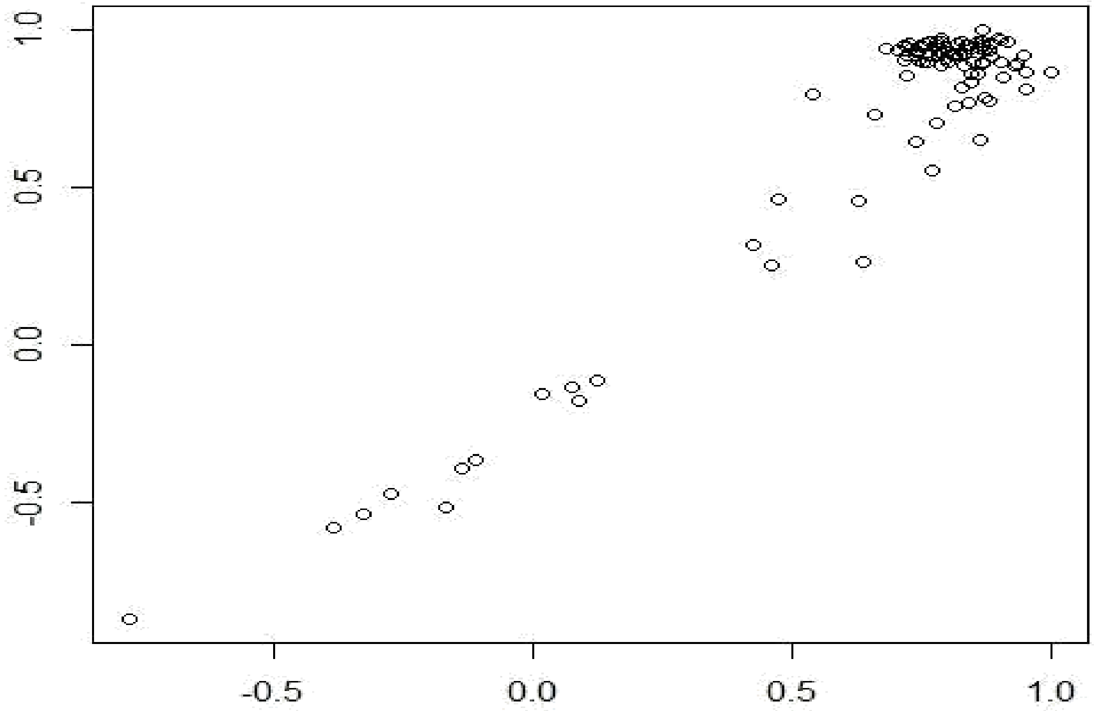


Fig 8: Correlation plot for the top 50 ranking genes. In this, Pearson correlation coefficient plot of genes with respect to each other is shown here, x- axis and y- axis are showing the genes.

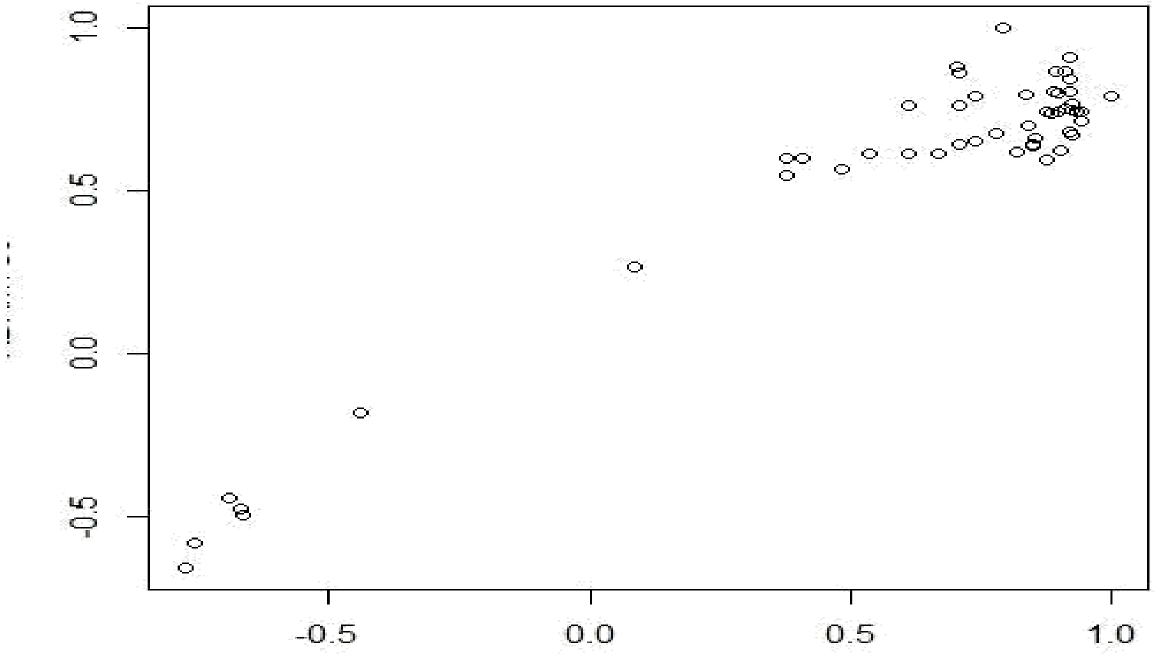


Fig 9: Correlation plot for the top 100 ranking genes. In this, Pearson correlation coefficient plot of genes with respect to each other is shown here, x- axis and y- axis are showing the genes.

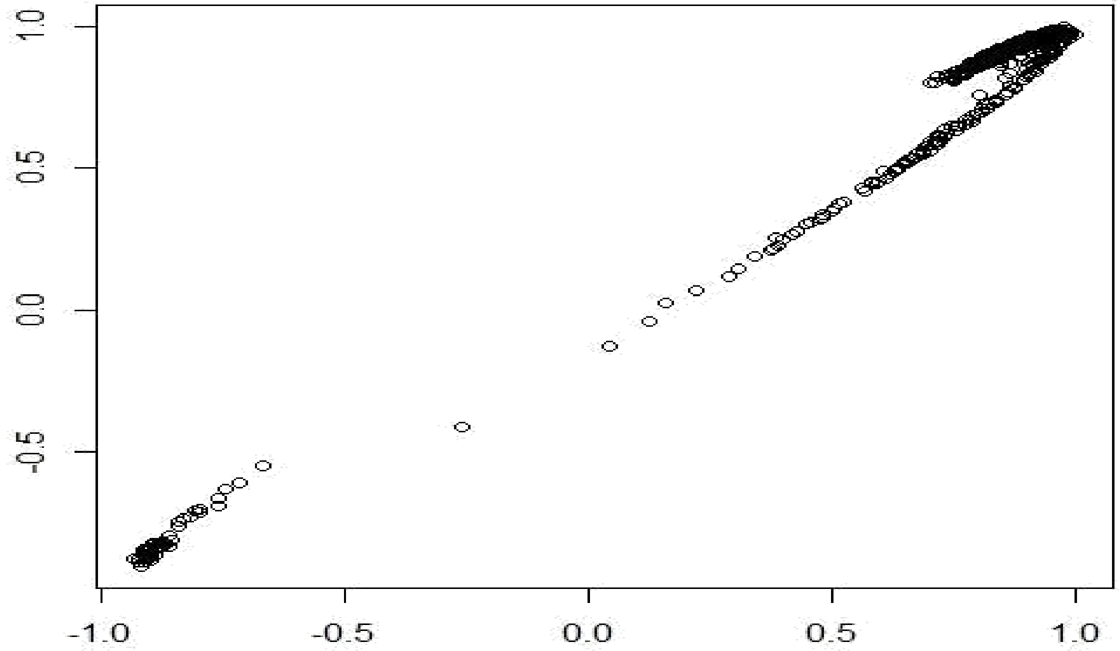


Fig 10: Correlation plot for the top 300 ranking genes. In this, Pearson correlation coefficient plot of genes with respect to each other is shown here, x- axis and y- axis are showing the genes.

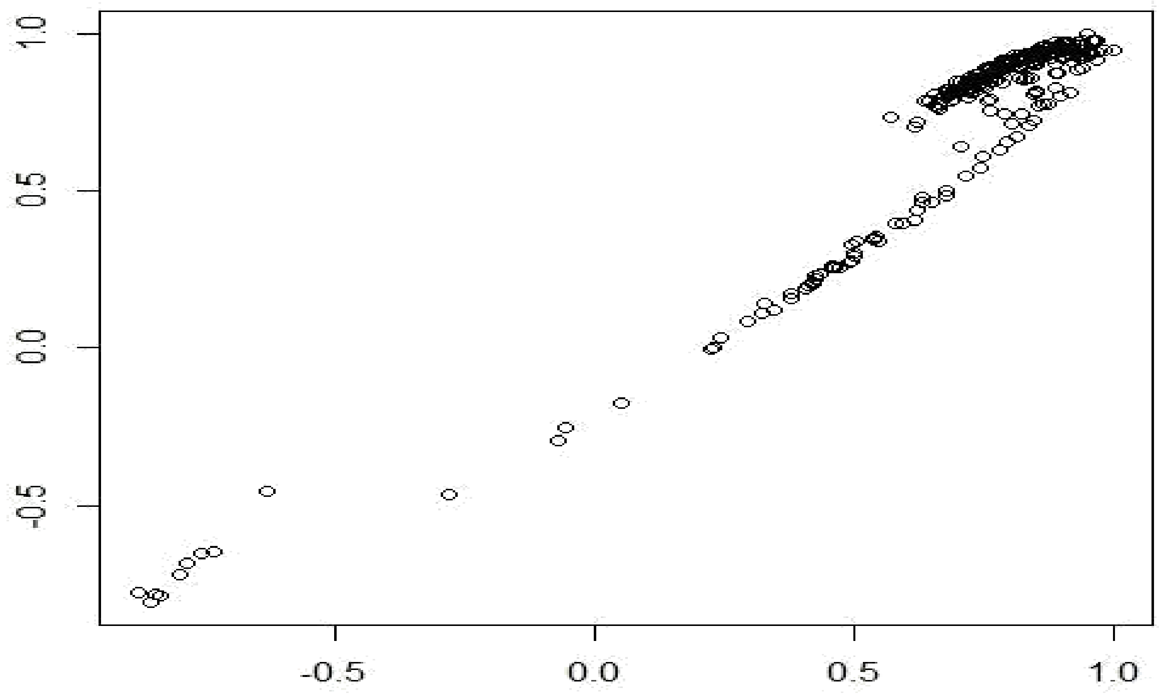


Fig 11: Correlation plot for the top 500 ranking genes. In this, Pearson correlation coefficient plot of genes with respect to each other is shown, x- axis and y- axis are showing the genes.

CLUSTER ANALYSIS

Clustering method makes graph which depicts more correlated genes in the same cluster and distant in different clusters. Results are displayed in the Figure 12, 13, 14, 15 and in Table 10, 11, 12, 13. In Table 10, 11, 12, and 13, it locates the cluster number of each gene for different values of k.

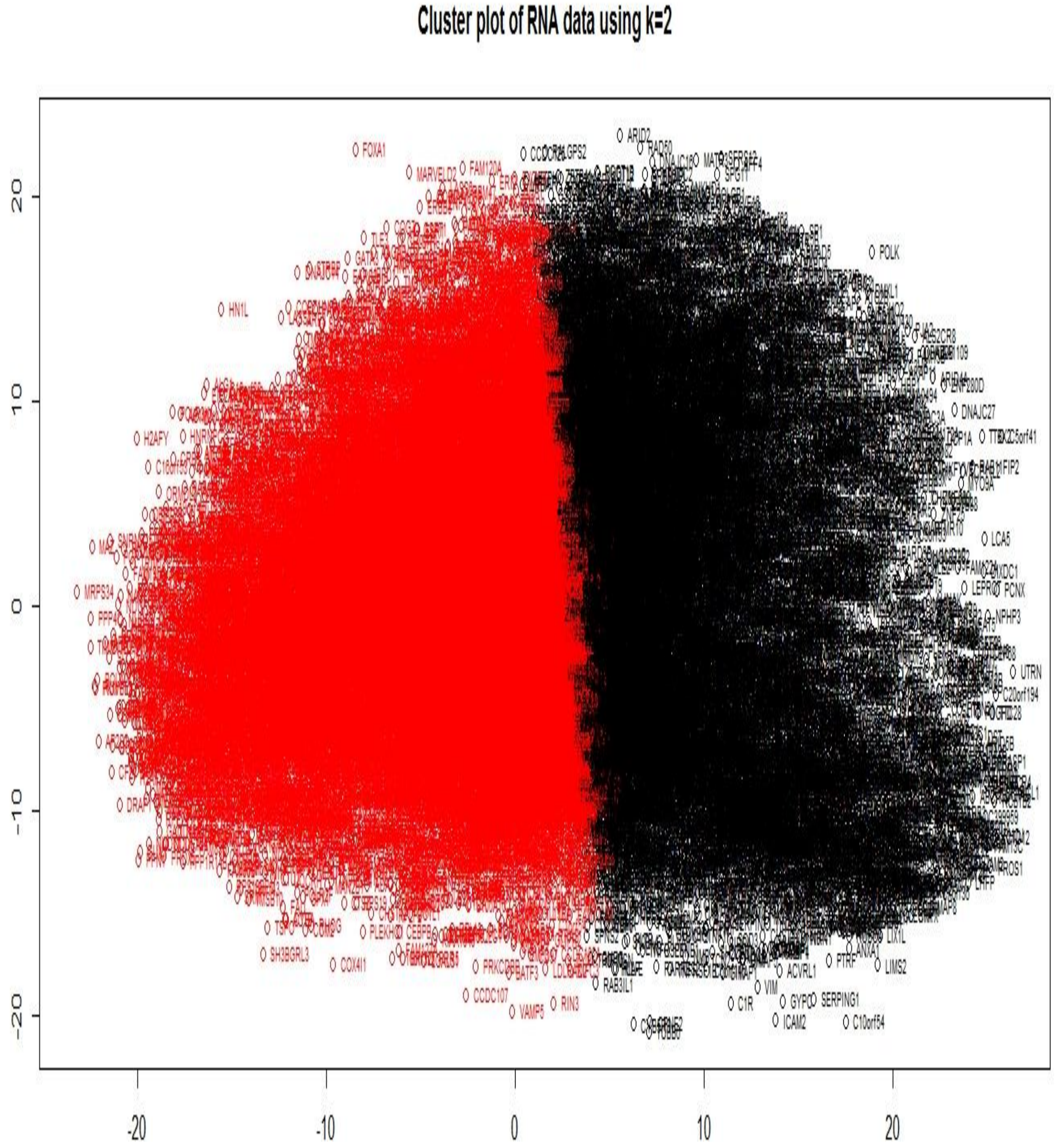


Fig 12: Cluster plot of gene expression data using k = 2. Black color in graph shows cluster 1 and red color shows cluster2.

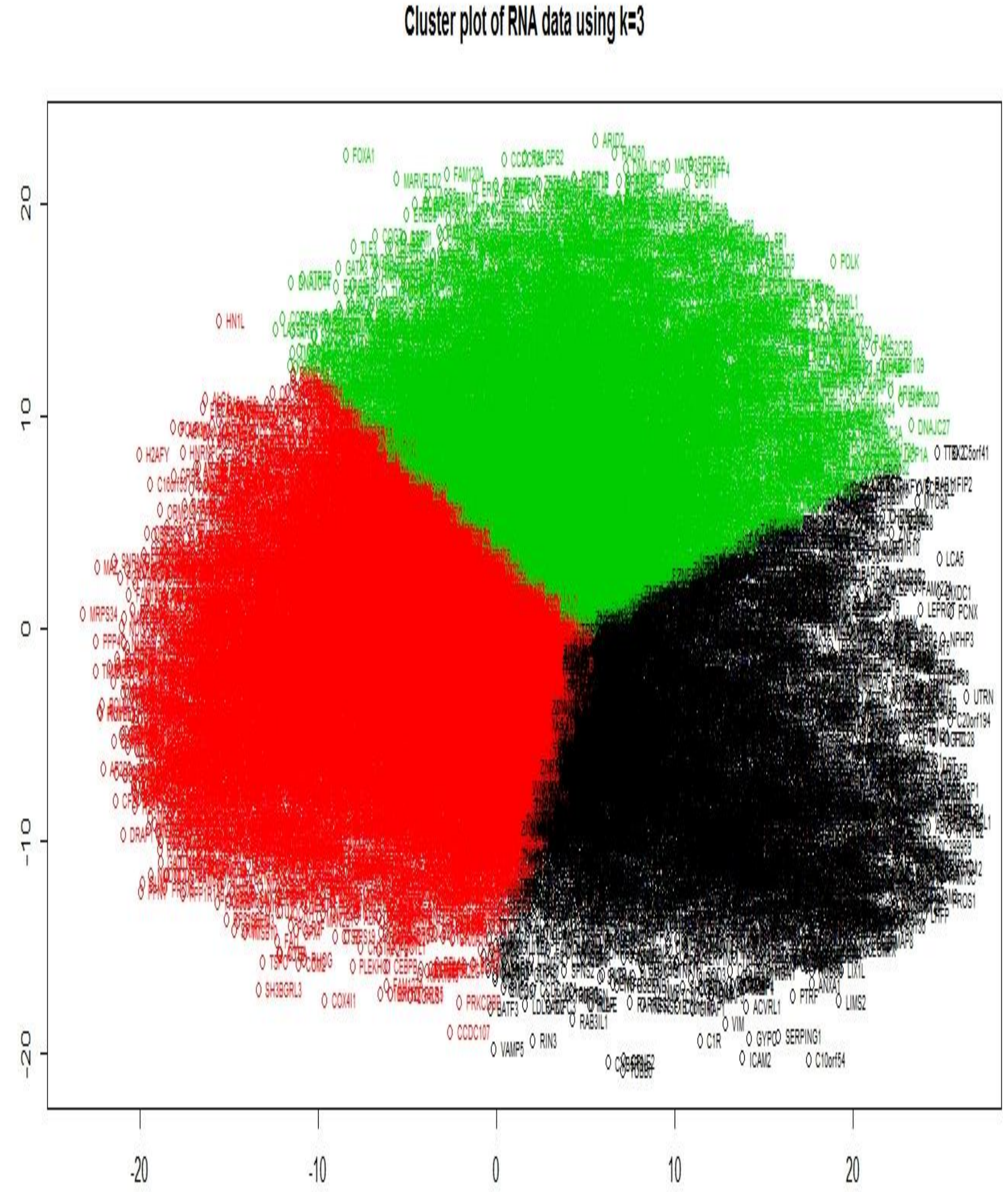


Fig 13: Cluster plot of gene expression data using k = 3. Black color in graph shows cluster 1 and red color shows cluster 2 and green color shows cluster 3

`

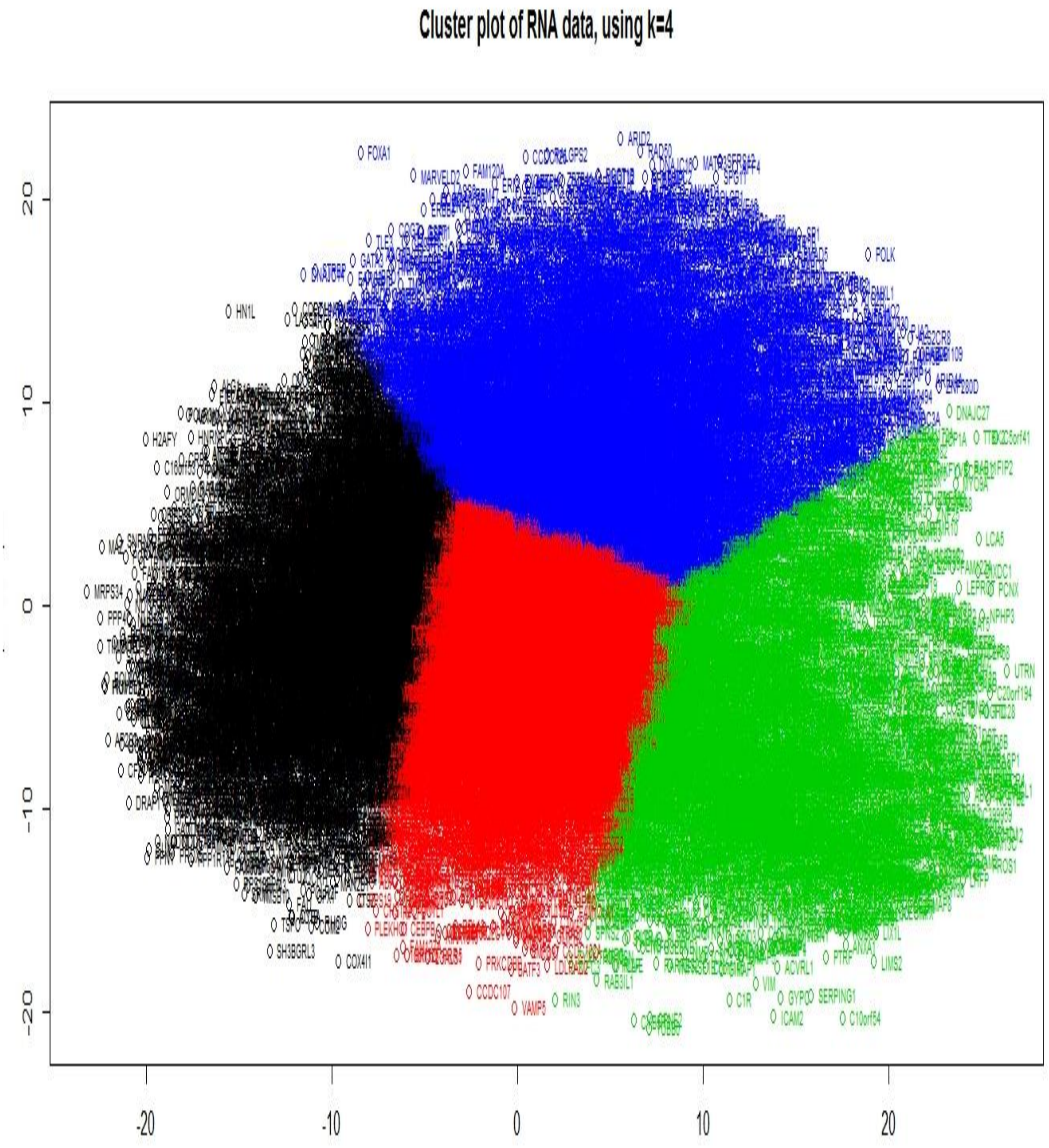


Fig 14: Cluster plot of gene expression data using k = 4. Black depicts cluster 1, red shows cluster 2, green is cluster 3 and blue shows cluster4.

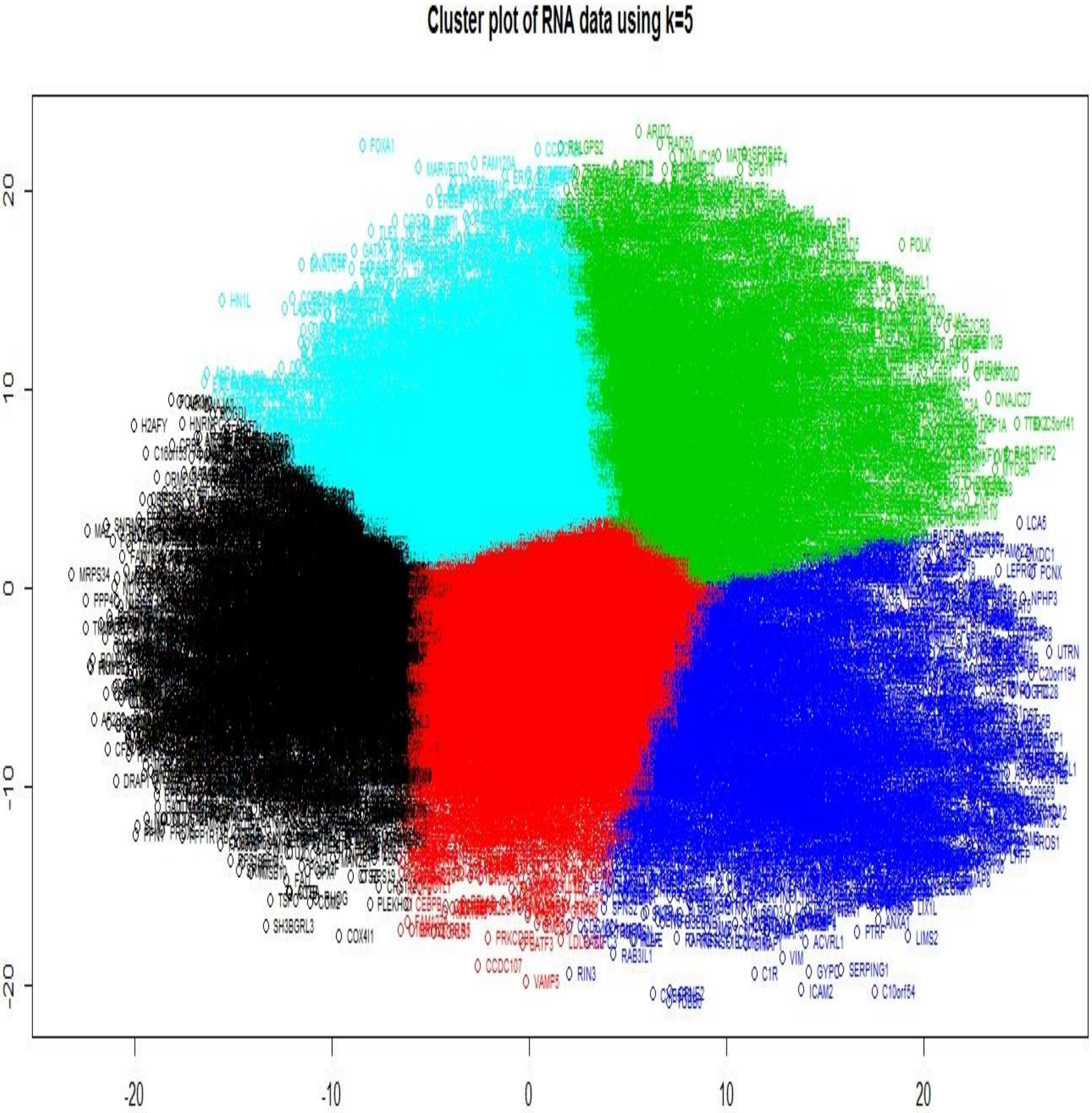


Fig 15: Cluster plot of gene expression data using k = 5. Black color shows cluster 1, red is cluster 2, green is cluster 3, dark blue is cluster 4, light blue shows cluster 5.

6. DISCUSSION AND FUTURE PROSPECTIVES

Recent decades have shown tremendous improvements in the area of high throughput sequencing which led to exponential growth of data. The efficient processing, storage and retrieval of such data becomes the fundamental requirement for interpretation of gene expression data, multidisciplinary approaches have become inevitable to deduce the underlying biological information. Genomics is the fundamental of cancer; large numbers of genes are involved in this, to study each and every gene is not an easy task. We need to develop a consensus approach to prioritize the gene and our study is conducted to deal with this problem.

RNA_Seq technology enables the measurement of expression level of thousands of genes and also provides knowledge to the researcher and gives insight to biological phenomena.

Various statistical softwares are available for data analysis, but we did our analysis in R-statistical software and any other software can also be used to analyse and compare the data to retrieve the information.

Gene Rank method considers some previous knowledge of gene connectivity and added information of its biological process and molecular function for its ranking procedure. It is the basis of Google's Page Rank algorithm (Morrison *et al*., 2005), so scientist should spend time to that gene which comes with high rank for further analytical advancements. This method improves the understanding of gene connections in the sample data.

In (E. Demidenko, 2015) studies, he has taken several microarray data sets and calculated the gene rank to show the complex gene connections in organs and organisms. They have used a priori knowledge (constant damping factor) for their calculation. Here we have calculated gene rank for Breast Cancer RNA_Seq gene expression data from TCGA with different damping factors, gives similar ranking of the genes with some difference in their values. GR approach and clustering shows close relation with respect to each other, so high rank genes are present in same cluster due to more correlated to each other.

In (Butte *et al*., 1999), they studied Pearson correlated coefficient which was performed by a program written in MATLAB to find the network of related variables in medical datasets. Various correlation measures are available to determine the closely related gene in gene expression data. Pearson correlation coefficient perceived as the proximity measures and no broad study has been conducted to analyse other correlation coefficient in our study. It tells about linear relationship of gene in respect to each other.

In order to get information contained in the gene expression data, the most used approach is making cluster of the genes, which gathers genes with similar expression pattern in same cluster and different in different clusters. A very easy way to discovering interesting gene is comparison of expression profile of differentially expressed genes (Brazma *et al*., 2000; 2003). The clustering result is displayed in graphic manner which represents both clustering and basal expression data of gene in a precise and coherent manner. Genes in same cluster shows that genes share some common cellular processes (Eisen *et al*., 1998). Co-expressed gene might be due to statistics capture similarity in shape. In their study, hierarchical clustering based on average linking method was used to compute dendogram. Our analysis was conducted using K-means algorithm to compute the clusters, this method is attributable to both simplicity and practicality and we found that genes within the same cluster sharing some cellular processes. Different algorithm based clustering method may also be used to analyse gene expression data.

In this study, we have considered sample data from diseased tumor, this approach can also be used to compare diseased, control and treated data which can provide helpful information of gene expression in different states of sample. These approaches may be used for a homogeneous sample or for comparison of gene connectivity among cases and controls.

In our analysis, we have taken three different damping factors in Gene rank calculation for diseased state. In future it can be used by taking more different damping factors for the gene rank calculation in different state of sampled data.

In our study, TCGA database has been used to collect the RNA_Seq gene expression data file to prioritize the gene, data from other databases may also be taken to analyse the gene expression for any diseases to get more useful information. We conducted these approaches using RPKM variable, other variable can also be used to analyse the data.

Ranking of gene can be helpful in making assumptions for the gene pathways generation and comparison of gene connectivity. To understand the genetic complexity, more studies are required.

The genetic information status may have advantages in-depth understanding of the diseases and in its treatment planning. So, the top ranking genes determined from these techniques may be useful to the researchers for further analysis in genomics. It may provide in-sight to the driver of mutation and can be beneficial in personalized medicine. Finding the genes from this way may provide potential target for therapeutic analysis in various diseases. It improves the ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease.

7. CONCLUSION

To analyse the gene expression data of Breast cancer, we collected RNA_Seq data available from TCGA for giving the preference to the genes in our study. We extracted the genes RPKM variable values (one of the normalization process in RNA_Seq) from that data.

During file preparation we came across that all the data files were not having completely similar genes. It shows some redundant gene, also the genes carrying zero values (did not express in the data sample) and some unidentified genes. To make the data in precise and coherent manner, we eliminated these anomalies and formed the consistent file.

In statistical analysis using R-software, to get meaningful information of gene we used Gene Rank, Pearson correlation coefficient, and Cluster techniques.

In Gene Rank, the file of genes with their gene rank value was obtained. The genes with higher position gets higher rank in file. Higher ranking gene also shows their involvement in other type of tumor. The density of genes at extreme position in graph is low.

In Pearson correlation coefficient, relation of genes with respect to each other gives the stability of genes expressing together in sample. The positive value shows if one gene expression is increasing, the other will also be increasing, and vice versa depicts the coexpression of genes. On the contrary, negative value shows if one is increasing other will be decreasing, and vice versa depicts if one is expressed then other is not. The values greater than 0.5 threshold shows more related genes with each other in the data, they may express coherently.

In clustering approach, every cluster contains various genes which shows some similarity pattern such as in case of cluster calculated using k=4, Cluster 1 having AURKB involved in mitotic spindle attachment to the centromere (Gisselsson *et al.*, 2008*)*, TUBA1 called tubulin α in structural molecule is related with cell division activity (Stotz *et al.*, 1999) and TPX2 is involved in a spindle assembly protein (Wittmann *et al.*, 2000). In Cluster 2, CDCA7 involves in cell division cycle associated with proliferation (Janicki *et al.*, 2011), MSH2 in DNA repairing (Schofield *et al.*, 2003), LBR provides interaction between lamina B and chromatin (Pyrpasopoulou *et al.*, 1996). Cluster 3, RBMS3 shows some cytoplasmic functions (Penkov *et al.*, 2000), LDB2 in regulating the cell migration (Storbeck *et al.*, 2009) and PLSCR4 involves in regulation of translocation of phospholipids (Phillippe *et al.*, 2006). In case of Cluster4, gene such as PJA2 responsible for proteasomal degradation (Yu *et al.*, 2002), KIA1109 involves in regulation of cell growth and differentiation (Dutertre *et al.*, 2010), POLK in DNA repair (Wood *et al.*, 2005) and BAZ2B (Jones *et al.*, 2000) in transcription regulation and in components of the remodelling complexes of chromatin. In the same way in all clustering approaches, separation of groups shares some pattern. Genes within the same cluster shows some similarity pattern for its clustering and also correlated to each other.

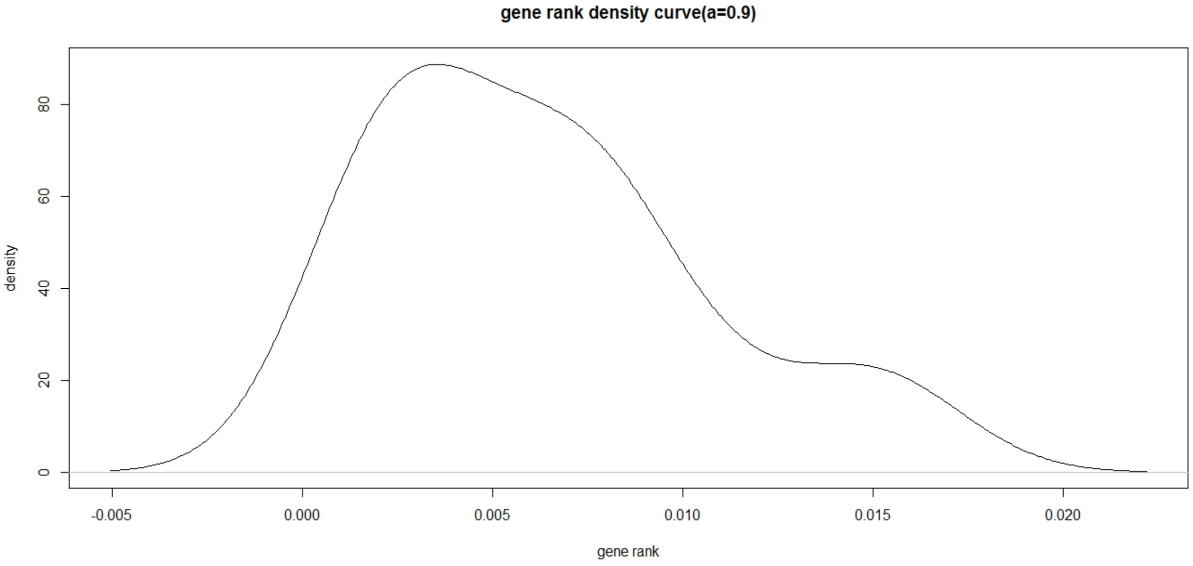


Fig 16: Gene rank density curve for the most susceptible gene in Breast Cancer.

In Figure 16, gene rank density curve for most susceptible gene in Breast cancer found by (De Jong *et al.*, 2002) has been shown, which tells that the higher ranking gene were present in low density region.

In Table 14, we have shown top 20 ranked genes from each cluster and analyse its Pearson correlation coefficient values with respect to genes, coloured cells shows the most correlated genes..

In Table 15, shows Pearson correlation coefficient value of the Breast cancer susceptible gene in our data. And Table 16 shows susceptible genes, rank value corresponding cluster number for a=0.80, 0.85 and 0.9

From all these techniques we came to know, that genes within the cluster, shows high rank and more correlated to each other against the other genes. So from this study, we found that results obtained from these methods can be used to predict gene priority.

8. REFERNCES

Anders, C; Carey, LA. (2008).Understanding and treating triple-negative breast cancer. Oncology (Williston Park). ***11***, 1233-9.

Bo, TH; Dysvik, B; Jonassen, I. (2004). LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research. ***32***, e34.

Bertucci; Francois. (2008). Reasons for breast cancer heterogeneity. Journal of Biology. ***7***, 6.

Brazma, A.; Vilo, J. (2000). Gene expression data analysis. FEBS letters. ***480(1)***, 17-24.

Brazma, A; Parkinson; Sarkans; U, Shojatalab ; Vilo., J; Abeygunawardena, N; Holloway, E, Kapushesky ; Kemmeren, P; Lara, GG; et al. (2003). Array Express–a public repository for microarray gene expression data at the EBI. Nucleic Acids Research. ***31***, 68–71.

Butte, A. J.; Kohane, I. S. (1999). Unsupervised knowledge discovery in medical databases using relevance networks. Proceedings of the AMIA Symposium. 711–715.

Carol A, Parise; Vincent, Caggiano. (2014). Breast Cancer Survival Defined by the ER/PR/HER2 Subtypes and a Surrogate Classification according to Tumor Grade and Immunohistochemical Biomarkers. Hindawi Publishing Corporation. Journal of Cancer. ***11***.

Chandrasekharappa, S.C.; Lach, F.P.; Kimble, D.C.; Kamat. A; Teer, J.K.; Donovan, F.X.;Flynn, E.; Sen, S.K.; Thongthip, S; Sanborn, E.; et al. (2013). Massively parallel sequencing, aCGH, and RNA-Seq technologies provide a comprehensive molecular diagnosis of Fanconi anemia. Blood. ***121***, e138-e148.

Chen, C; Li, Z; Yang, Y; Xiang, T; Song, W; Liu, S. (2015). Microarray Expression Profiling of Dysregulated Long Non-Coding RNAs in Triple-Negative Breast Cancer. Cancer Biol Ther. ***21***.

Churpek, J.E.; Walsh, T.; Zheng, Y.; Moton, Z.; Thornton, A.M.; Lee, M.K.; Olopade, O.I. (2015). Inherited predisposition to breast cancer among African American women. Breast Cancer Research and Treatment. ***149***, 31–39.

Cloonan, N; et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods. ***5***, 613–619.

De Jong, M. M.; Nolte, I. M.; Te Meerman, G. J.; Van der Graaf, W. T. A.; Oosterwijk, J. C.; Kleibeuker, J. H.; De Vries, E. G. E. (2002). Genes other than BRCA1 and BRCA2 involved in breast cancer susceptibility. Journal of medical genetics. ***39(4)***, 225-242.

Dutertre, M.; Lacroix-Triki, M.; Driouch, K.; de la Grange, P.; Gratadou, L.; Beck, S.; Auboeuf, D. (2010). Exon-based clustering of murine breast tumor transcriptomes reveals alternative exons whose expression is associated with metastasis. Cancer research, ***70(3)***, 896-905.

D'haeseleer, P.; Wen, X.; Fuhrman, S.; Somogyi, R. (1998). Mining the Gene Expression Matrix: Inferring Gene Relationships From Large Scale Gene Expression Data. Information Processing in Cells and Tissues. 203–212.

Eisen; Michael, B.; Spellman; Paul, T.; Brown; Patrick, O.; Botstein, David. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA. ***95(25)***, 14863–14868.

ENCODE Data Matrix. Retrieved 2013-07-28.

Eugene Demidenko. (2015). Microarray enriched gene rank. Bio Data Mining. ***8***, 2.

F. Galton. (1877). Typical laws of heredity. Nature. ***15 (492–495)***, 512–514.

F. Galton. (1885). The British Association: Section II, Anthropology: Opening address by Francis Galton, F.R.S., etc., President of the Anthropological Institute, President of the Section. Nature. ***32 (830)***, 507–510.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute of Great Britain and Ireland. ***15***, 246–263.

Galton; Pearson; the Peas.(2001). A Brief History of Linear Regression for Statistics. Journal of Statistics Education. ***9***, 3.

Gisselsson, D.; Hakanson, U.; Stoller, P.; Marti, D.; Jin, Y.; Rosengren; A. H.; Panagopoulos, I. (2008). When the genome plays dice: circumvention of the spindle assembly checkpoint and near-random chromosome segregation in multipolar cancer cell mitoses. PLoS One. ***3(4)***, e1871.

Hartuv, E.; Schmitt, A.; Lange, J.; Meier-Ewert, S.; Lehrach, H; Shamir, R. (1999). An algorithm for clustering cDNAs for gene expression analysis using short oligonucleotide fingerprints. In Proceedings of the Third International Symposium on Computational Biology. (RECOMB***99***). ACM Press, New York. 188–197.

Heikkinen, T.; Kärkkäinen, H.; Aaltonen, K.; Milne, R.L.; Heikkilä, P.; Aittomäki, K.; Nevanlinna, H. (2009). The breast cancer susceptibility mutation PALB2 1592delT is associated with an aggressive tumor phenotype. Clinical Cancer Research. ***15***, 3214–3222.

Hill, J.T.; Demarest, B.L.; Bisgrove, B.W.; Gorsi, B.; Su ,Y.C.; Yo, H.J.(2013). MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. Genome Res. ***23***, 687-697.

Janicki, P.; Boeuf, S.; Steck, E.; Egermann, M.; Kasten, P.; Richter, W. (2011). Prediction of in vivo bone forming potency of bone marrow-derived human mesenchymal stem cells. Eur Cell Mater. ***21***, 488-507.

Jiang, Daxi; Chun, Tang; Aidong, Zhang. (2004). Cluster analysis for gene expression data: A survey. Knowledge and data Engineering. ***16(11)***, 1370-1386.

Julie L, Morrison; Rainer, Breitling; Desmond J, Higham; David R, Gilbert. (2005). GeneRank: Using search engine technology for the analysis of microarray experiments. BMC Bioinformatics. ***6***, 233.

Jones, M. H.; Hamana, N.; Nezu, J. I.; Shimane, M. (2000). A novel family of bromodomain genes. Genomics. ***63(1)***, 40-45.

Karl Pearson. (1985). Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London. ***58***, 240-242.

Khatoon, Z; Figler, B; Zhang, H; Cheng, F. (2014). Introduction to RNA-Seq and its applications to drug discovery and development. ***75(5)***, 324-30.

Kumar, Dhiraj; Santanu Kumar, Rath; Abhishek, Pandey. (2009). Gene Expression Analysis Using Clustering. IEEE. 1 – 4.

Lehmann, BD; Pietenpol, JA ; Tan, AR. (2015). Triple-negative breast cancer: molecular subtypes and new targets for therapy. ***35***, e31-9.

M., Bittner; P., Meltzer; Y., Chen; Y., Jiang; E., Seftor; M., Hendrix; M., Radmacher; R., Simon; Z., Yakhini; A., Ben-Dor; N., Sampas; E., Dougherty; E., Wang; F., Marincola; C., Gooden; J., Lueders; A. ,Glatfelter; P., Pollock; J.,Carpten; E., Gillanders; D., Leja; K., Dietrich; C., Beaudry; M., Berens; D., Alberts; V., Sondak; N., Hayward;J., Trent. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. ***406***, 536-540.

Miller, A.C.; Obholzer, N.D.; Shah, A.N.; Megason, S.G.; Moens, C.B. (2013). RNA-Seq based mapping and candidate identification of mutations from forward genetic screens. Genome Res. ***23***, 679-686.

Morin, R; et al. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques. ***45***, 81–94.

Mortazavi , A; Williams, BA; McCue, K; Schaeffer, L; Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Method. ***5***, 21-8.

Muhammad Rukunuddin, Ghalib; Rittwika, Ghosh; Priti, Sasmal; Udisha, Pande. (2013). Microarray gene expression data analysis using enhanced K-means clustering method. IJAET. ***5***, 373-380.

Mwenifumbo, J. C.; Marra, M.A. (2013). Cancer genome-sequencing study design. Nat. Rev. Genet. ***14(5)***, 321-332.

Nagalakshmi , U; et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. **320**, 1344–1349.
O'Brien, K.M.; Cole, S.R.; Engel, L.S.; Bensen, J.T.; Poole, C.; Herring, A.H.; Millikan, R.C. (2014). Breast cancer subtypes and previously established genetic risk factors: A bayesian approach. Cancer Epidemiology, Biomarkers, and Prevention. ***23***, 84–97.

Page, L; Brin, S; Motwani, R; Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. Tech rep Stanford Digital Library Technologies Project.

Parmigiani, G.; Garrett, E. S.; Irizarry; R. A.; Zeger, S. L. (2003). The analysis of gene expression data: an overview of methods and software. Springer New York. 1-45.

Penkov, D.; Ni, R.; Else, C.; Piñol-Roma, S.; Ramirez, F.; Tanaka, S. (2000). Cloning of a human gene closely related to the genes coding for the c-myc single-strand binding proteins. Gene. ***243(1)***, 27-36.

Phillippe, M.; Bradley, D. F.; Ji, H.; Oppenheimer, K. H.; Chien, E. K. (2006). Phospholipid scramblase isoform expression in pregnant rat uterus. Journal of the Society for Gynecologic Investigation. ***13(7)***, 497-501.

Pyrpasopoulou, A.; Meier, J.; Maison, C.; Simos, G.; Georgatos, S. D. (1996). The lamin B receptor (LBR) provides essential chromatin docking sites at the nuclear envelope. The EMBO journal. ***15(24)***, 7108.

Schofield, M. J.; Hsieh, P. (2003). DNA MISMATCH REPAIR: Molecular Mechanisms and Biological Function. Annual Reviews in Microbiology. ***57(1)***, 579-608.

Shah ,S.P.; Kobel, M.; Senz, J.; Morin, R.D.; Clarke, B.A.; Wiegand, K.C.; Leung, G; Zayed, A.; Mehl, E.; Kalloger, S.E.; et al. (2009). Mutation of FOXL2 in granulosa-cell tumors of the ovary. N. Engl. J. Med. ***360***, 2719-2729.

Somogyi, R.; Wen, X.; Ma, W.; Barker, J.L. (1995). Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. Journal of Neuroscience. ***15(4)***, 2575-2591.

Spellman; P.T.; Sherlock, G; et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell. ***9***, 3273–3297.

Storbeck, C. J.; Wagner, S.; O'Reilly, P.; McKay, M.; Parks, R. J.; Westphal, H.; Sabourin, L. A. (2009). The Ldb1 and Ldb2 transcriptional cofactors interact with the Ste20-like kinase SLK and regulate cell migration. Molecular biology of the cell. ***20(19)***, 4174-4182.

Stotz; Henrik, U.; Sharon, R. Long. (1999). Expression of the pea (Pisum sativum L.) α-tubulin gene TubA1 is correlated with cell division activity. Plant molecular biology. ***41(5)***, 601-614.

Stigler, Stephen M. (1989). Francis Galton's Account of the Invention of Correlation. Statistical Science. ***4 (2)***, 73–79.

The Cancer Genome Atlas- Data Portal. Retrieved 2013-07-28.

Troyanskaya, O; Cantor, M; Sherlock, G; Brown, P; Hastie, T; Tibshirani, R; Botstein, D; Altman, RB. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics. ***17***, 520-525.

U, Alon; N., Barkai; D. A., Notterman; K., Gish; S., Ybarra; D. ,Mack; A.J.,Levine. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. ***96(12)***, 6745–675.

Viale; G. (2012). The current state of breast cancer classification. Annals of Oncology. ***23***, 207-210.

Vogelstein, B; Kinzler, KW. (2004). Cancer genes and the pathways they control. Nat Med. ***10(8)***, 789-799.

Wang, Z; Gerstein, M.; Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Gener. ***10***, 57-63.

Wen, X.; Fuhrman, S.; Michaels, G.S.; Carr, D.B.; Smith, S.; Barker, J.L.;Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. U.S.A. ***95(1)***, 334–339.

Winter, C; Kristiansen, G; kersting, S; Roy, J; Aust ,D;Knosel, T; Rummele, P, et al.(2012). Google goes cancer: Improving outcome prediction for cancer patients by network- based ranking of marker genes. PLOS Comput Biol. ***8,*** e1002511.

Wittmann, T.; Wilm, M.; Karsenti, E.; Vernos, I. (2000). TPX2, A novel xenopus MAP involved in spindle pole organization. The Journal of cell biology. ***149(7)***, 1405-1418.

Wu, V. S.; Kanaya, N.; Lo, C.; Mortimer, J.; Chen, S. (2015). From bench to bedside: What do we know about hormone receptor-positive and human epidermal growth factor receptor 2-positive breast cancer?. Journal of Steroid Biochemistry and Molecular Biology.

Wood, R. D.; Mitchell, M.; Lindahl, T. (2005). Human DNA repair genes, 2005. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. ***577(1)***, 275-283.

Xu, X.; Zhu, K.; Liu, F.; Wang, Y.; Shen, J.; Jin, J.; Wang, Z.; Chen, L.; J.; Xu, M.(2013). Identification of somatic mutations in human prostate cancer by RNA-Seq. Gene. ***519***, 343-347.

Yeung, KY; Haynor, DR; Ruzzo,WL. (2001). Validating clustering for gene expression data. Bioinformatics. ***17(4)***, 309-18.

Yu, P.; Chen, Y.; Tagle, D.A.; Cai, T. (2002). PJA1, encoding a RING-H2 finger ubiquitin ligase, is a novel human X chromosome gene abundantly expressed in brain. Genomics. ***79(6)***, 869-874.

Zainab, Khatoon ; Bryan, Figler ; Hui, Zhang ; Feng, Cheng.(2014). Introduction to RNA-Seq and its Applications to Drug Discovery and Development. Drug Discovery. ***75***, 324–330.

Zhang, B; Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology. ***4***, Article 17.

Zuber, V; Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. Bioinformatics. ***25***, 2700-7.

9. APPENDIX

METHODOLOGY CODE IN R

RETREIVING RPKM VARIABLE FROM DATA FILE

```
setwd("drivename:/foldername")

p1=filename[,c(gene column number, RPKM column number)]

row.names(p1)=p1$gene

head(p1)

p1$gene=NULL

write.csv(p1, "RPKMfilename.csv")
```

DATA FILE MERGING

```
setwd("drivename:/foldername")

l=list.files()

a1=read.csv(l[1],header=T,sep=",")

for(i in 2:length(l)){l1=read.csv(l[i],header=T,sep=",");

 colnames(l1)[1]="gene";

colnames(l1)[2]=l[i];a1=merge(a1,l1,by="gene")}

write.csv(a1, “mergedfilename.csv”)
```

TRANSPOSE CODE FOR MERGING FILE

```
k1=read.csv("mergedfilename.csv",sep=",",header=T,row.names=1)

k2=as.matrix(k1)

k3=t(k2)

write.csv(k3,"transposemergedfilename.csv")
```

CORRELATION COEFFICIENT

```
a=read.csv("transposemergedfilename.csv",header=T, sep=",", row.names=1)

d=cor(a)

write.csv(d,"correlationfilename.csv")
```

CORRELATION PLOT

```
X= read.csv(“merge file name”, header=T, sep=”,”, row.names=1)

Y=cor(X)

 plot(Y)
```

GENE RANK

```
B= read.csv("mergedfilename.csv", header=T, sep=",", row.names=1)

 x.bar=rowMeans(B)

 Bsub.mean=B-x.bar

 sdB=sqrt(rowSums(Bsub.mean^2))

 Z=(1/sdB)*Bsub.mean

eps=0.0001;maxit=10

 a=0.9

 n=nrow(Z)

 m=ncol(Z)

 sumR2=rep(0,n)

 for(i in 1:m)

 for(j in 1:m)

 {

 tij=sum(Z[,i]*Z[,j])

 sumR2=sumR2+tij*Z[,i]*Z[,j]
 }
 s=sumR2/sqrt(sum(sumR2^2))
 for(it in 1:maxit)
 {
 tR2s.fast=rep(0,n)
 for(i in 1:m)
 for(j in 1:m)
 {
 lij=sum(Z[,i]*Z[,j]*s/sumR2)
 tR2s.fast=tR2s.fast+lij*Z[,i]*Z[,j]
 }
 s.new=(1-a)/n*sum(s)+a*tR2s.fast
 s.new=s.new/sqrt(sum(s.new^2))
 adiff=max(abs(s-s.new))
 if(adiff<eps) break
 s=s.new
 }
Write.csv(s, “generankfilename.csv”)
```

GENE RANK DENSITY CURVE

```
h= read.csv("generank filename.csv", header=T, sep=",")
 fs= density(h$rank)
 plot(fs, main="gene rank density curve(a=0.8)", xlab="gene rank", ylab="density")
```

CLUSTERING

```
op1=Z

prc=prcomp(op1,scale=T,center=T)

head(prc$x)

pc.comp <- prc$x

pc.comp1 <- -1*pc.comp[,1]

head(pc.comp1)

pc.comp2 <- -1*pc.comp[,2]

X <- cbind(pc.comp1, pc.comp2)

cl <- kmeans(X,4)

plot(pc.comp1, pc.comp2,col=cl$cluster,main="Cluster PCA plot of RNA data")

points(cl$centers, pch=16)

l=cl$cluster

cl$cluster

Y<- cbind(pc.comp1, pc.comp2,l)

colnames(Y)[3]="Type"

Z=as.data.frame(Y)

text(Z$pc.comp1, Z$pc.comp2, row.names(Z), cex=0.6, pos=4, col=cl$cluster)

write.csv(l,”cluster.csv”)
```